FACULTE DE MEDECINE
COCHIN PORT-ROYAL
UNIVERSITE PARIS V

ANNEE 1971

THESE

N° 132

POUR LE

DOCTORAT EN MEDECINE

(Diplôme d'Université)

PAR

Joseph FERRETTI

Né le 30 mai 1941 à New York

PRESENTEE ET SOUTENUE PUBLIQUEMENT LE : . . . . . .

L'ETAT ACTUEL DE LA DOCUMENTATION MEDICALE AUTOMATIQUE

Président : M. H. PEQUIGNOT, Professeur

EDITIONS MEDICALES ET UNIVERSITAIRES

40, rue Pascal - PARIS XIIIe

FACULTE DE MEDECINE
COCHIN PORT-ROYAL
UNIVERSITE PARIS V

ANNEE 1971

THESE

N°

POUR LE

DOCTORAT EN MEDECINE

(Diplôme d'Université)

PAR

Joseph FERRETTI

Né le 30 mai 1941 à New York

PRESENTEE ET SOUTENUE PUBLIQUEMENT LE : . . . . . .

L'ETAT ACTUEL DE LA DOCUMENTATION MEDICALE AUTOMATIQUE

Président : M. H. PEQUIGNOT, Professeur

EDITIONS MEDICALES ET UNIVERSITAIRES

40, rue Pascal - PARIS XIIIe

UNIV.
LEIDEN
BIBL.

I dedicate this work to my parents,

my wife Chantal and her family.

Without their help I would never have completed my medical studies.

I dedicate this work to my parents,

my wife Chantal and her family.

Without their help I would never have completed
my medical studies.

A Monsieur le Professeur H. PEQUIGNOT

Qui m'a fait l'honneur d'accepter la présidence de cette thèse.

Qu'il reçoive ici l'expression de mes sentiments respectueux.

A Monsieur le Professeur M. TEQUIGNOT

Qui m'a fait l'honneur d'accepter la présidence
de cette thèse.

Qu'il reçoive ici l'expression de mes sentiments
respectueux.

Au Docteur Harold VANDENBURGH

Mon ami, programmeur itinérant,

qui m'a suggéré le sujet de cette thèse.

Qu'il reçoive ici mes plus grands remerciements.

Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41550028_0010

P L A N

INTRODUCTION

PREMIERE PARTIE : La Documentation Automatique Générale

A) Introduction
B) La Chaîne Documentaire
C) Le Langage Naturel et les Langages Documentaires
D) L'Analyse Automatique des Documents
E) L'Organisation des Fichiers et les Stratégies de Recherche
F) La Diffusion de l'Information Documentaire
G) L'Evaluation des Systèmes Documentaires
H) Le Coût de la Documentation Automatique
I) Les Systèmes Question-Réponse Automatiques
J) Conclusion

DEUXIEME PARTIE : Systèmes de Documentation Automatique Opérationnels et Spécifiques de la Médecine.

A) Introduction
B) MEDLARS
C) Excerpta Medica
D) Le Système du Cox Coronary Heart Institute
E) SABIR
F) DIMDI

G) FAIR
H) L'OMS et la Documentation Automatique
I) Conclusion

TROISIEME PARTIE : Services Documentaires Automatiques Non Spécifiquement Médicaux

A) Introduction
B) Biological Abstracts, BASIC, et BioResearch Index
C) Chemical Abstracts, Chemical Titles et Chemical Abstracts Condensates
D) Science Information Exchange (S I E)
E) Automatic Subject Citation Alert (A S C A)
F) Science Citation Index (S C I)
G) Current Contents, Life Sciences. Permuterm Subject Index (P S I)
H) Conclusion

QUATRIEME PARTIE : La Banque de Données

A) Introduction
B) Systèmes de Recherches de Données
C) La Banque de Données en Enzymologie
D) Conclusion

CONCLUSION

BIBLIOGRAPHIE :
I - BIBLIOGRAPHIE SELECTIVE

II - REFERENCES SUPPLEMENTAIRES

INTRODUCTION

Le but de cette thèse est de faire le point sur la documentation automatique et particulièrement sur ses applications médicales. Par système documentaire on entend un ensemble de règles et de procédures concernant l'enregistrement et la recherche des informations scientifiques (en anglais, "Information storage and retrieval"). L'automatisation peut intervenir à différents moments dans le fonctionnement d'un système de documentation. On appelle documentation automatique, toute application de procédés mécaniques ou de méthodes de calcul (au sens large du mot) pour l'analyse et/ou la recherche d'informations, particulières (27).

Discipline à peine née, la documentation automatique a fait un progrès immense depuis 10 ans. Cette évolution sera explorée sous son aspect général dans la première partie de la thèse. Dans la deuxième partie, des exemples précis seront exposés, en insistant sur le projet MEDLARS. La troisième partie présentera des exemples de services documentaires non spécifiques de la médecine mais utilisables par elle. A part, et dans la quatrième partie, le problème des banques de données sera étudié, en prenant comme exemple la banque de données de l'INSERM appliquée à l'enzymologie.

La bibliographie occupe une place importante dans ce travail. Une sélection de références, que je crois représentative de la littérature, est présentée dans une première bibliographie. La seconde, beaucoup plus courte, est recommandée par l'équipe Salton et contient une documentation récente sur le projet SMART.

L'Index Medicus étant la source classique et souvent la plus accessible aux médecins, il m'a semblé utile de compléter mes recherches bibliographiques par une fouille de l'Index Medicus de 1970 à 1960. Les rubriques suivantes (Medical subject Headings) ont été surtout examinées : Abstracting and Indexing ; Documentation ou Bibliography ; Information Retrieval Systems.

Grâce à un document plus adapté à l'informatique que l'Index Medicus, j'ai pu explorer la littérature plus ancienne (156). Il s'agit d'une bibliographie comportant 6100 références concernant l'informatique publiées entre 1946 et 1963. Ce "Computer Literature Bibliography" de W. W. Youden représente une expérience de recherche rétrospective des informations (information retrieval) pour déterminer la valeur d'un index KWIC dont il sera question dans la première partie de cette thèse.

Cinq critères ont servi dans le choix des références de la première bibliographie : commodité, importance, facilité d'accès, fréquence de citation et actualité. Par "commodité", j'entends un document surtout utile pour la rédaction de cette thèse ; par "actualité", des documents publiés dans les 10 dernières années. On peut critiquer la "fréquence de citation" comme moyen de choix car les auteurs les plus cités ne sont pas forcément les plus importants (90). Certains auteurs importants, Cleverdon par exemple, ne sont pas cités parce que je n'avais pas un accès direct à leurs publications. Notons que cette bibliographie ne représente qu'une minime partie de la littérature concernant la documentation automatique accumulée depuis seulement 10 ans (13, 43, 71).

Au total, il y a 158 citations dans la bibliographie sélective, dont environ un tiers provient de journaux médicaux. Le tableau suivant montre le type de journal le plus souvent cité.

Tableau I : Journaux les plus cités

Journal	Nombre de Citations
American Documentation	15
Science	14
Bull. Med. Libr. Ass.	12
J. ACM	11
Commun. ACM	10
JAMA	10
La Presse Médicale	9
Le Concours Médicale	5
La Revue d'Informatique Médicale	5

PREMIERE PARTIE

LA DOCUMENTATION AUTOMATIQUE GENERALE

A) Introduction

"La masse des données médicales acquises, dont la connaissance est nécessaire aux soins des malades, dépasse les capacités de préhension et de mémoire du médecin. Nul médecin ne plus plus connaître la médecine" (56) L'information médicale avec ses techniques documentaires automatiques offre à ce problème une solution réaliste. A ce propos, la définition de l'informatique du Professeur Grémy cadre bien avec les buts de la documentation automatique. "L'informatique est l'ensemble des moyens et méthodes, intellectuels et techniques, qui permettent de gérer rationnellement l'information. Par gestion de l'information, il faut entendre : le recueil, le stockage, l'exploitation et la diffusion de l'information". Sans entrer dans les discussions des cybernéticiens sur le sens du mot "information", disons simplement qu'information égale "toute donnée représentable en ordinateur" (26, 42, 96).

B) La Chaîne Documentaire

Les activités d'un centre de documentation peuvent être représentées schématiquement ainsi :

Documents $\xrightarrow[\text{documentaire}]{\text{analyse}}$ (1) Mémoire (2) $\xleftarrow[\text{documentaire}]{\text{recherche}}$ Demandes d'Informations

(1) = mise en mémoire

(2) = consultation de mémoire

Les documents arrivant au centre sont analysés avant leur mise en mémoire. L'analyse documentaire est l'opération par laquelle on tend à exprimer le contenu d'un document, sous une forme différente de la forme originale. Il ne s'agit donc pas de conserver la totalité de l'information contenue dans le document original, mais seulement cette partie de l'information qui permettra à un chercheur de retrouver ce document. Ceci implique une indexation des documents. L'indexation exprime le contenu des documents au moyen de descripteurs (mots-clés), avec ou sans indications syntaxiques; elle détermine la qualité d'une documentation. Une bonne indexation doit tenir compte autant de la façon dont les questions seront posées au service de documentation que du contenu du document (27, 79, 149).

L'arrivée d'une demande d'information au centre déclenche une recherche documentaire. La recherche documentaire consiste à sélectionner, dans un ensemble de documents, les informations répondant à une demande particulière. On distingue dans cette procédure :

- l'analyse de la demande (chaque question est formulée à l'aide de mots-clés concernant le sujet intéressant le demandeur) ;
- la consultation de la mémoire (la sélection des documents pertinents repose sur la coïncidence entre les mots-clés de la question et les mots-clés associés aux documents) ;
- la sortie des réponses (une liste est établie indiquant sous des formes variables les références de ces documents).

Le système de recherche le plus efficace est celui dans lequel cette comparaison aboutit à la sélection des documents ayant un intérêt maximum pour le demandeur, et cela pour un coût total minimum de recherche (27, 150).

C) Le Langage Naturel et les Langages Documentaires

Le langage naturel est le langage couramment utilisé pour décrire ce que nous observons. Ce langage à une double structure : sémantique et syntaxique. La structure sémantique organise les rapports de signification des concepts et des mots qui les expriment. Elle reconnaît des rapports de voisinage sémantique. La structure syntaxique relie ces éléments significatifs en une phrase cohérente, elle-même significative. Ces structures permettent au langage naturel d'exprimer toute notion nouvelle par une combinaison des éléments sémantiques et syntaxiques préexistants (54, 58, 94, 103).

Le "Chaos sémantique" règne dans les langues naturelles. Elles sont toutes irrégulières dans leur grammaire, leur syntaxe, leur vocabulaire. Les mots ont plusieurs sens, qui varient avec le contexte, avec la place dans la phrase, avec le style. Ces éléments ont déjà fait échouer les recherches sur la traduction électronique des langues (15, 24, 31). En médecine l'imprécision du langage médical constitue un obstacle majeur pour établir des communications correctes avec les ordinateurs. Il faut réduire le nombre de termes utilisés en médecine, éliminer ou classer les synonymes, répertorier les éponymes, préciser les éléments constitutifs de la définition des maladies, se référer aux notions admises à l'étranger. En un mot, il faut uniformiser et normaliser la nomenclature médicale. Le recours à l'ordinateur fournit une occasion impérieuse de réviser définitions et classifications ; chaque terme employé doit avoir une signification précise et les diverses catégories doivent être mutuellement exclusives (8, 15, 22, 53).

Le langage documentaire est un langage artificiel utilisé dans l'indexation des documents et la recherche documentaire. Ce langage peut-être considéré sous les deux aspects traditionnels : sémantique et syntaxique. L'aspect sémantique (ou lexique documentaire) est l'ensemble des termes (mots-clés, descripteurs) utilisés pour désigner les concepts, dans les re-

présentations de documents appartenant à un champ donné. La syntaxe documentaire (grammaire) est l'ensemble des procédés utilisés dans ce langage pour exprimer les relations logiques entre mots-clés, dans les représentations de documents (27).

Le langage documentaire (ou langage d'indexation) est un langage "normalisé" destiné à réduire les ambiguités du langage naturel. Ainsi les concepts contenus dans les documents et dans les questions, traduits dans un langage normalisé, auront plus de chance de coïncider.

On peut distinguer trois types de langages documentaires.

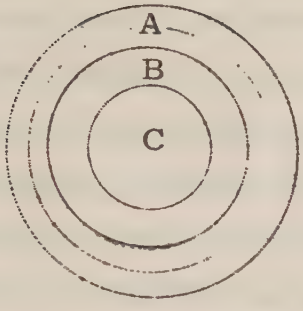
1 - Systèmes classificatoires : Ce sont des systèmes de classement logique, à structure hiérarchisée procédant par inclusion de classes. On a donc une série de classes emboîtées, où les classes les plus fines sont sous les classes les plus générales ; il y a une place et une seule pour la notion représentée. La Classification Décimale Universelle (C D U), qui prend pour base une division en 10 à chaque niveau, est un exemple courant de ce type d'organisation documentaire. La nature rigoureuse de la classification permet d'utiliser les possibilités de tri des machines. Pourtant ces systèmes ne donnent pas entière satisfaction. La difficulté du codage augmente au fur et à mesure que le volume des documents impose des subdivisions de plus en plus fines ; une notion nouvelle peut remettre en cause toute la classification antérieure. Les chercheurs se méfient du caractère universel du C D U, et préfèrent adopter la documentation automatique à un domaine particulier (126, 129, 149, 151, 154, 157).

2 - Systèmes à structure combinatoire : Ce sont des systèmes à juxtaposition de descripteurs opérant par intersection de classes. Au lieu de se fonder sur une organisation "à priori" du langage, on part du document dont on relève les mots importants (mots-clés, descripteurs). Les

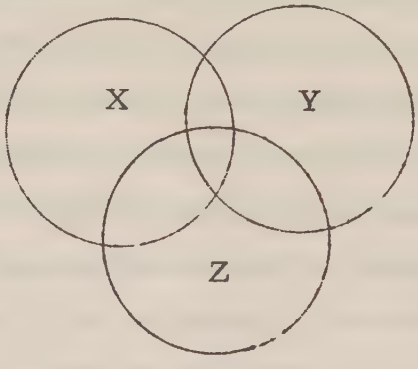
notions sont rangées par ordre alphabétique ; les notions voisines sont ainsi dispersées. L'ordre en est entièrement arbitraire du point de vue logique, mais on a un accès direct à la notion cherchée (151, 154, 157).

On peut opposer la structure hiérarchisée rigide du premier système d'organisation linguistique à la grande souplesse des systèmes combinatoires (154) :

Types de structures de langage d'indexation



Structure hiérarchisée
(inclusion de classes)



Structure combinatoire
(intersection de classes)

L'utilisation d'un système à structure combinatoire nécessite le recours à un dictionnaire organisé de telle manière que chacun des descripteurs qui y figurent fournisse une liste de mots synonymes, hiérarchiquement liés ou sémantiquement voisins. On appelle ce dictionnaire "thésaurus" qu'on peut définir plus précisément comme étant un dictionnaire d'équivalences entre les mots ou expressions d'une langue naturelle et les descripteurs d'un langage documentaire (27, 64, 80, 87).

En médecine le thésaurus est la somme des connaissances au sein d'une spécialité, un dictionnaire des symptômes, signes syndromes et maladies, contenant en outre, l'étude de leurs relations d'équivalence, d'asso-

ciation ou d'incomptabilité, de leur fréquence réciproque au sein d'une population. Si le thésaurus offre aux cliniciens un vocabulaire de référence, on assiste depuis quelques années, à une profilération des thésaurus, souvent compétitifs et incompatibles entre eux, qui ne fera qu'augmenter les difficultés de communication (21, 72, 97).

3 - Systèmes à syntaxe : Les systèmes à structure combinatoire ne permettent pas de rendre compte des relations existant entre les notions contenues dans un document. Il peut y avoir, dans un document donné, soit absence de relation (absorption par la muqueuse buccale de médicaments : barbituriques ou enzymes) soit relation réversible (action des enzymes sur le métabolisme des barbituriques). L'introduction d'une syntaxe permet de réduire le nombre de documents non pertinents apparus lors de la recherche. Il est aussi possible de représenter les liaisons entre les termes en définissant un jeu de relations logiques ou sémantiques dans un domaine déterminé. Le SYNTOL est un exemple de ce type de langage documentaire à syntaxe. On y rangera aussi les grilles d'analyse utilisées dans la banque de données de l'INSERM. Il s'agit d'un langage à syntaxe préformée jouant le rôle d'un questionnaire et assurant ainsi à l'analyse un certain degré de cohérence (39, I23, I28, I35, I57).

Ainsi, les buts d'un langage documentaire sont : économiser le nombre des symboles employés dans le fichier descripteur ;standardiser les descriptions des sujets, afin d'accroître la probalilité de coïncidence entre celles de l'analyste et du demandeur ; augmenter la probabilité de retrouver, pour une question, tous les documents pertinents, et aucun document non pertinent ; fournir des références spécifiques ou des réponses générales dans la mesure des besoins des utilisateurs (150).

L'efficacité de performance d'un système de documentation dépend essentiellement de l'exhaustivité (recall) de l'indexation et de la spécificité (précision) du langage documentaire. Ces facteurs interviennent aussi bien

dans la stratégie de recherche des documents que dans l'analyse documentaire (76). Cette question sera plus développée dans la discussion sur l'évaluation des systèmes documentaires.

Pour conclure, voici deux exemples de langages documentaires. Le premier et plus vieux est né en 1926. Son auteur, Hans Selye, lui a donné le nom de Symbolic Shorthand System (SSS) pour la Physiologie et la Médecine. Le Dr. Selye a appliqué son système à l'indexation d'environ 700 000 documents sur le stress, la calciphylaxie, et l'endocrinologie à l'Institut de Médecine et de Chirurgie expérimentales, Université de Montréal. Il s'agit d'une classification analytico-synthétique comprenant 20 classes principales basées sur les grands systèmes de l'organisme (Système Nerveux, Système Respiratoire, etc...). A l'intérieur de chaque classe, il existe une organisation à facettes permettant de considérer un document simultanément de plusieurs façons. Ainsi la classe Système Gastro-intestinal est subdivisée selon les facettes :

- Structure (les parties du système) ;
- Intervention (chirurgie) ;
- Fonction (fonctions naturelles du système) ;
- Produits spécifiques (hormones, secrétions) ;
- Maladies.

Le SSS est énumératif en partie. Il énumère des étiquettes pour des sujets simples de la littérature. Il est également synthétique ; il existe des règles permettant la synthèse des sujets élémentaires pour exprimer des sujets extrêmement complexes (76, 132).

Exemples :
Adr = la glande surrénale
Psy = le psychisme
Hyp-x = hypophysectomie
ACTH = adrenocorticotrophic hormone
TX = Thyroxine

Donc, Adr ← Psy veut dire "l'effet d'un stimulus psychique sur la glande surrénale".

Ou encore, Adr ← Hyp-x + ACTH + TX exprime "l'effet sur la glande surrénale de hypophysectomie en conjonction avec l'ACTH et la thyroxine".

Le deuxième langage artificiel est le SYNTOL (Syntagmatic Organisation Language) de J. C. Gardin. Le SYNTOL est un ensemble de règles logico-linguistiques visant différentes manières de représenter les informations trouvées dans la documentation scientifique, pour faciliter l'enregistrement et la manipulation sur calculateurs électroniques.

Le SYNTOL complète l'organisation paradigmatique (organisation analytique ou verticale) des thesaurus par une organisation syntagmatique (organisation synthétique ou horizontale) exprimant donc les relations entre les termes. Une des caractéristiques du système est que les expressions admises, pour représenter le contenu des documents, se réduisent à des chaînes du "syntagmes" élémentaires. On entend par syntagme tout couple de mots-clés lié par une relation logique.

La grammaire du SYNTOL comprend quatre relations logiques : coordinative, consécutive, associative, prédicative. Les mots SYNTOL sont rangés en catégories (prédicats, entités, états, actions). Ces mots sont choisis pour représenter les idées principales contenues dans un document, puis ils sont identifiés en catégories par une consultation de dictionnaire. Ensuite les relations sont déterminées entre syntagmes, et la direction de chaque relation (par exemple de l'entité à l'état, ou de l'entité à l'action) est spécifiée par des règles préétablies. Chaque syntagme avec sa relation correspondante peut-être représenté de la manière suivante :

A = action
E = entité
S = état

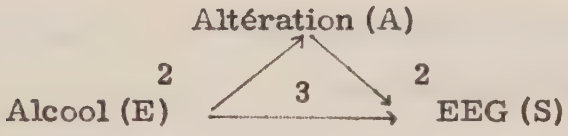
2 = relation associative
3 = relation consécutive

Donc, Altération (A) $\xrightarrow{2}$ EEG (S) veut dire "altération de l'E E G".

Alcool (E) $\xrightarrow{2}$ Altération (A) signifie "l'alcool altère quelque chose".

Alcool (E) $\xrightarrow{3}$ EEG (S) égale "l'alcool modifie l'E E G".

Le rassemblement de ces syntagmes donne comme résultat :



ce qui veut dire "altération de l'EEG due à l'alcool"
(Références : 7, 14, 27, 94, 103, 122, 149).

Ces procédures d'analyse syntaxique prennent tout leur intérêt à la fois dans l'automatisation des recherches rétrospectives et dans l'analyse automatique de contenu de documents.

D) l'Analyse Automatique des Documents

"Les systèmes de recherche documentaire sont construits et utilisés en fonction d'une série de transformations, grâce auxquelles les documents originaux sont enregistrés de telle manière qu'ils puissent être fournis en réponse à une question" (150). Le tableau suivant montre ces transformations dont on a déjà parlé en présentant la chaîne documentaire.

Tableau II : Transformations Documentaires.

Transformations d'Enregistrement	Transformations de Recherche
- adressage et enregistrement du texte original	- mise de la question sous forme lisible par l'ordinateur
- mise sous forme lisible par l'ordinateur	- formalisation de la question
- formulation de la spécification bibliographique	- remplissage de la fiche de recherche
- affectation d'une adresse d'enregistrement	- sélection de fiches intéressantes dans le fichier
- construction et sélection d'énoncés informatifs	- fournitures des spécifications et adresses
- formulation des descripteurs standards des sujets	- localisation des documents originaux
- remplissage des fiches et classement dans le fichier	- fourniture des documents au demandeur

L'opération d'enregistrement et de recherche documentaire peut-être divisée en quatre parties principales :

a - formulation des entrées de documents ;

b - construction du fichier de descripteurs ;

c - adressage, enregistrement, localisation et fourniture de documents ;

d - recherche dans le fichier de descripteurs et fourniture des résultats.

C'est l'ensemble de ces opérations qu'on entend automatiser. Les trois opérations (b, c, d) n'impliquent que des tâches de routine : transmission, enregistrement, tri, copie, comparaison, mise en fichier. Les tâches de l'opération (a) mise en catalogue, édition, extraction, formalisation, nécessitent un exercice de jugement (150). Donc, si l'on veut automatiser cette phase (l'indexation) du système il faut, soit communiquer une culture à la machine, soit lui donner les moyens de se la forger elle-même, par des méthodes d'apprentissage (27). Voilà l'ampleur dupproblème.

Quel est l'intérêt de la mécanisation de l'indexation ?

Un des résultats de l'explosion de l'information est l'augmentation de la redondance et de la répétition des efforts humains. Cette duplication du travail déjà fait est souvent très coûteuse. Mais il est parfois moins cher de répéter une expérience que de rechercher tous les documents en rapport avec celle-ci. Une pseudo-loi, attribuée à Mooers, dit qu'un système d'information tend à ne pas être utilisé quand il est plus "douloureux" pour l'utilisateur d'avoir l'information que de ne pas l'avoir. Donc, la clé de la recherche documentaire se trouve dans une indexation efficace. Or "des dizaines de milliers de spécialistes consacrent une part notable de leur temps à exprimer le contenu de documents scientifiques toujours plus nombreux, en vue de faciliter les recherches rétrospectives ultérieures ; le recrutement et la formation d'analystes compétents, pour cette tâche, sont de plus en plus difficiles, et il est naturel que l'on cherche à contourner l'obstacle par la mécanisation de celle-ci" (14, 138).

Pour représenter le contenu informationnel d'un document on peut utiliser :

- soit la méthode des mots-clés qui permet d'extraire les éléments les plus caractéristiques du texte ; une dizaine de mots-clés suffit en général pour représenter un article (cette méthode nécessite la lecture de l'article par un spécialiste capable de le comprendre - spécialiste de la question traitée ou spécialiste de la documentation) ; mais les analyses faites par des personnes différentes, utilisant le même thesaurus, peuvent aboutir à des mots-clés différents ;

- soit des méthodes automatiques se servant, quand il existe, du résumé fourni par l'auteur du document ; on demande à l'ordinateur de faire lui-même l'analyse de ce résumé ; le thesaurus se constitue au fur et à mesure et automatiquement à partir des mots utilisés dans le résumé (32, 38, 39, 78, 123, 149).

H. P. Luhn a étudié un mode de sélection automatique des mots, expressions et phrases. Il postule que "les mots les plus significatifs d'un texte sont ceux qui y figurent les plus fréquemment, mis à part les mots vides, tels que : de, le, un, par, sur, dans, pour, ceci, est, sont, etc...". On doit donc compter la fréquence d'apparition de chacun des mots significatifs du texte et sélectionner, comme représentant le sujet traité, les mots les plus fréquents. Dans une deuxième étape, on détermine le nombre de fois où des mots ayant la fréquence la plus forte se suivent dans une même phrase ; ces paires de mots peuvent caractériser plus étroitement le contenu du document. La dernière étape consiste à sélectionner les phrases du texte où la concentration des mots ou paires de mots précédents est la plus élevée ; ces phrases forment le résumé (plutôt l'extrait) de l'article (30, 40, 41, 88, 92, 138, 142, 150).

Exemple du Procédé de Luhn

Titre de l'article : Des expériences suggèrent un nouveau mode de traitement des crises cardiaques.

Liste de fréquence :

a) Mots simples =

hormones	14	médical	6
femelle	11	homme	6
effet	8	sexe	6
animal	7	traitement	6
cholestérol	7	crises	5
coeur	7	sang	5
médicaments	7	suggère	5

b) Paire de mots =

hormones, femelles	5	effet, sexe	3
crise, cardiaque	5	sang, cholestérol	2
hormones, sexuelles	3	effets, hormones	2
sexe, féminin	3		

c) Phrase retenue : Le résultat constitue au moins une étape, vers la découverte de composés qui agiront comme des hormones femelles, en abaissant le taux de cholestérol et dans le sang, chez les patients même souffrants de crise cardiaque, sans produire d'effets féminisants.

Les systèmes d'analyse automatique à base de construction de listes de fréquence pour l'ordinateur nécessitent la mise du texte sous forme lisible par la machine. Cette opération est très coûteuse actuellement mais les progrès dans le domaine de la lecture optique et de la reconnaissance des formes permettent d'espérer que la solution de ce problème est proche. Plus sérieux est le problème linguistique. Le calcul des fréquences des mots néglige le contexte dans lequel ils se trouvent. Ainsi les travaux plus récents s'occupent, d'une part du calcul de coefficients d'association

statistique entre termes et documents, d'autre part de la formation d'une variété d'algorithmes de classement automatique des termes et des documents. Rappelons qu'on entend par algorithme, un procédé basé sur la décomposition d'une tâche générale en un ensemble de tâches élémentaires reliées les unes aux autres selon un enchaînement strict.
(Références : 17, 33, 37, 38, 81, 83, 105, 106, 107, 121, 139).

Une solution à l'automatisation de l'indexation est apportée par les "KWIC index" (Key Word In Content) ou index de mots-clés dans le contexte. Le "KWIC index" est entièrement construit par l'ordinateur, à l'exception du choix des mots considérés comme non significatifs, il a l'avantage de fournir l'environnement du mot retenu et d'accélérer la recherche.

En pratique, l'index KWIC est limité aux titres ; il s'agit donc d'un KWIT (Key Word In Title) index. Dans cet index chaque titre est cité sous tous les mots significatifs qu'il contient. Ces mots-clés sont classés par ordre alphabétique avec le titre complet placé de telle façon que les mêmes mots-clés de plusieurs titres apparaissent l'un sous l'autre au milieu de la page :

Exemple : La cuisine au vin
Le vin du pays
Un bon vin blanc

Chaque titre s'accompagne d'une référence très abrégée permettant l'identification de l'article dans la partie bibliographique de l'index.

Le "Computer Literature Bibliography" 1946-1963 de W. W. Youden fournit un excellent exemple d'index KWIC. Ce document contient plus de 6 100 citations de la littérature concernant l'informatique publiée entre 1946 et 1963. Il comporte trois parties :

- une partie bibliographique qui cite les titres et les auteurs de tous les articles dans 2 journaux, 2 livres et plus de 100 "proceedings" ;

- un index de titres permutés (index KWIC) qui permet de retrouver un article si une partie de son titre est connue ou de trouver tous les articles dont les titres contiennent un mot ou une phrase particulière ;

- un index des auteurs.

Une variante de l'index KWIC, est le KWOC (Key Word Out of Context) index, où les mots-clés sont isolés de leur contexte et imprimés dans la marge gauche de la page , suivis du titre dans l'ordre normal.

Les avantages d'un index KWIC sont multiples :

- il est d'une préparation simple et peu coûteuse ;
- il utilise peu de mots et n'a pas besoin d'une analyse avant le stockage ;
- les délais de publication sont courts ;
- la liste de références sert de point de départ dans la recherche d'un sujet dans un grand nombre d'articles.

Ses inconvénients sont les suivants :

- il est peu précis, long et peu agréable à lire ;
- le titre peut ne pas refléter le contenu d'un article ;
- l'utilisation du langage naturel favorise des ambiguïtés ;
- il est encombrant.

(Références : 47, 66, 89, 91, 112, 148, 156).

Dans le domaine de l'automatisation complète d'un système de documentation, le système SMART de Salton semble être le plus avancé. Ce système expérimental, développé entre 1962 et 1965, fonctionne actuellement aux Universités Harvard et Cornell aux Etats-Unis. Les ordinateurs IBM 7094 et IBM 360/65 sont utilisés. Les documents et les demandes sont acceptés en langage naturel.

Plusieurs facilités pour l'analyse documentaire sont incorporées dans le système.

1 - la méthode dite suffixe "s" et un dictionnaire des racines des mots (null thesaurus)

2 - un dictionnaire de synonymes ou thesaurus régulier

3 - un arrangement hiérarchique des concepts inclus dans le thesaurus montrant les rapports "parents", "fils" ou "frères" entre les concepts

4 - méthodes d'association statistique

5 - méthodes d'analyse syntaxique

6 - méthodes de reconnaissance statistique des phrases

7 - un système dictionnaire destiné à réviser les multiples dictionnaires inclus dans le système

8 - un système de classement automatique des documents

9 - un système permettant la participation de l'utilisateur (user feedback system)

Les documents enregistrés et les demandes sont traités dans le système SMART sans analyse manuelle auparavant. Plusieurs combinaisons de méthodes automatiques d'analyse de contenu sont utilisées,et les documents qui coïncident le mieux avec les demandes sont extraits du fichier. Plus précisément un coefficient de corrélation est calculé pour indiquer le degré de similarité entre chaque document et chaque demande. Les documents sont ensuite rangés par ordre décroissant des coefficients de corrélation.

Le système SMART permet d'évaluer l'efficacité de plusieurs méthodes de traitement en comparant les résultats obtenus par plusieurs recherches, ceci est réalisé en traitant la même demande plusieurs fois dans le même fichier, tandis qu'un changement de procédure d'analyse est fait entre chaque traitement. En comparant la performance de la formulation de recherche sous différentes conditions de traitement, il est possible de déterminer l'efficacité relative des différentes méthodes d'analyse. (Références : 30, 33, 75, 113, 121, 122, 124, 127, 130, 139).

E) L'Organisation des Fichiers et les Stratégies de Recherche

Dans l'analyse de l'information, une grande variété de méthodes est utilisée :

- méthodes à base de consultation de dictionnaire destinées à remplacer les mots ou les expressions du texte par des numéros de classe ;

- méthodes statistiques utilisant les caractéristiques des pairos de mots pour former des associations de termes et de documents ;

- méthodes d'analyse structurelle utilisant les relations syntaxiques des mots.

Ces procédures d'analyse sont destinées à remplacer un document donné ou une demande par un vecteur "d'identificateurs de contenu". Ces identificateurs de contenu (termes, concepts, mots-clés) sont ensuite utilisés pendant la recherche dans le fichier pour distinguer les éléments recherchés des éléments non-recherchés (121).

Le fichier d'enregistrement (ou de recherche) est un élément essentiel de tout système de documentation : il fournit un mécanisme par lequel les descriptions des documents peuvent être comparées aux descriptions des demandes. Un document est "retrouvé" quand sa description coïncide avec celle de la demande (76).

Il existe plusieurs modes d'organisation des fichiers :

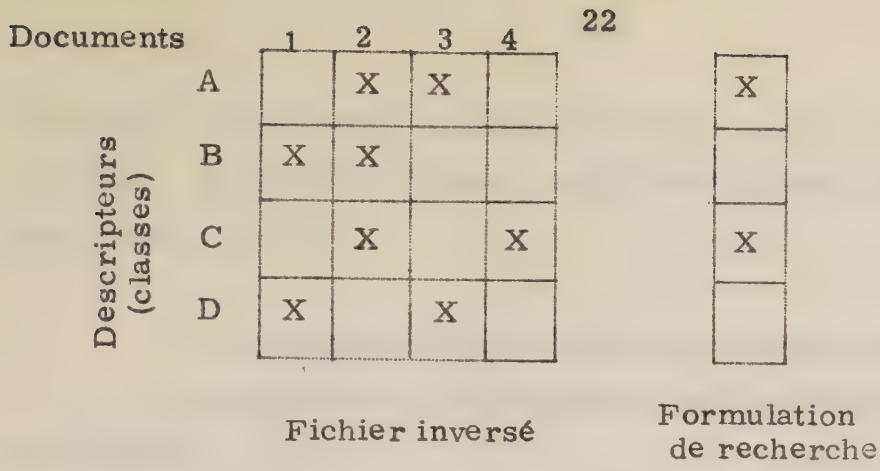
- fichier direct : les références bibliographiques des documents sont rangées suivant un ordre séquentiel quelconque, numérique ou alphabétique ; chaque référence est suivie de l'ensemble des descripteurs (mots-clés) ;

- fichier inversé : chaque descripteur est suivi d'une indication permettant l'identification des articles qui le comportent ;

- fichier mixte : un fichier inversé plus un fichier direct ; une recherche se fait en deux étapes ;

- fichier "multilist" : un procédé de chaînage est utilisé ; chaque descripteur du fichier direct est suivi de l'indication du prochain document où il se trouve (60, 73, 123, 140).

Voici un exemple de fichier inversé plus une formulation de recherche. Le fichier est représenté sous forme de "matrice documentaire".



La formulation de recherche (colonne de droite) représente une demande de documents inclus à la fois dans la classe A et C. L'opération de recherche consiste à comparer la formulation de recherche et la matrice documentaire. On voit que seul le document "2" cofncide avec la formulation donnée, même s'il est également inclus dans la classe B (76).

Le mode d'accès a l'ordinateur est déterminé par le type d'organe périphérique utilisé pour stocker les informations. Si le support matériel du fichier est la banqe magnétique, une recherche simple se fait de façon séquentielle, en lisant fiche par fiche le contenu de la bande ; chaque fiche correspond à un espace déterminé sur la bande. Pour répondre à une question, il faut donc dérouler toutes les bandes du fichier. Pour améliorer les performances et pour des raisons d'économie, on présente en même temps des références correspondant à plusieurs demandes (Batch Processing ou traitement par paquets) (149).

Si on bénéficie de mémoires à accès aléatoire (disques, tambours), mémoires où l'information est retrouvée à partir d'une adresse, il est possible d'utiliser l'accès direct. Ce mode d'accès à plusieurs avantages ; il réduit les temps d'occupation machine (dont le coût) ; le fichier inversé peut-être employé, une interaction entre l'utilisateur et la machine est possible ; il permet le traitement d'une question à la fois. En effet, l'utilisateur peut dialoguer avec l'ordinateur au cours de la recherche (mode conversationnel). Il travaille en "temps réel", c'est-à-dire que le système fonctionne de

telle façon que la réponse suit de très près la réception d'informations extérieures. Sa question, souvent ambigüe au départ, peut ainsi être précisée peu à peu (149, 157).

F) La Diffusion de l'Information Documentaire

Un système de documentation automatique peut rendre trois types de service :

- la fourniture de catalogues, de bulletins et d'index ;
- les recherches rétrospectives ;
- la diffusion personnalisée de l'information.

Les catalogues, les bulletins et les index peuvent être établis par l'ordinateur. Les bulletins peuvent être signalétiques ou analytiques. Il existe des index par mots-clés comme le KWIC index.

Les recherches rétrospectives permettent la constitution d'une bibliographie correspondant étroitement à une question posée. L'ordinateur récupère les références en comparant les données de la question avec les mots-clés stockés dans le fichier.

Peut-être le plus intéressant de ces services est la dissémination sélective de l'information de Luhn (en anglais : Selective Dissemination of Information ou S. D. I.). Il s'agit d'un service personnalisé qui offre aux individus et aux groupes la possibilité de recevoir, à mesure de leur arrivée, des informations sur les documents susceptibles de les intéresser. Pour cela l'abonné du service doit définir d'abord, en se servant de descripteurs, un profil correspondant à ses besoins. Ce "profil d'intérêt" peut-être considéré comme une question permanente. L'interrogation du fichier est limitée aux documents nouveaux. L'ordinateur compare donc périodiquement les descripteurs de ces questions permanentes avec ceux des documents nouvellement parvenus au centre de documentations. Autrement dit, dans un sys-

tème SDI, l'arrivée d'un document déclenche une recherche d'un utilisateur intéressé, tandis que dans un système de recherche rétrospective c'est l'arrivée d'une question qui précipite la recherche de documents.

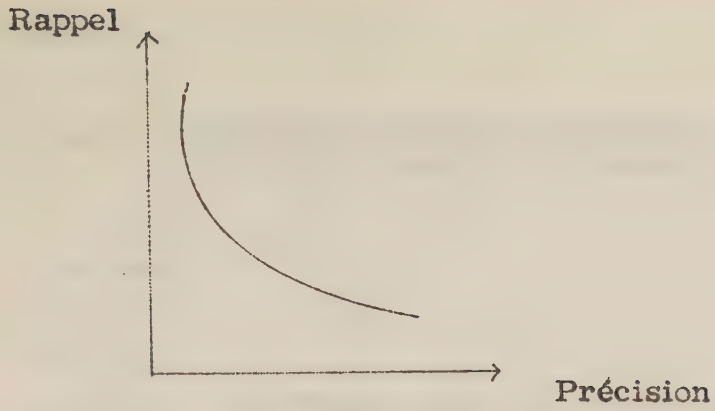
La qualité de ce service peut-être améliorée et constamment mise à jour grâce à la pondération des termes du profil d'intérêt. Le profil est modifié d'une fourniture à l'autre jusqu'à ce que les envois répondent aux besoins exacts du demandeur. Avec la connaissance de la pertinence ou non des documents signalés, l'ordinateur peut établir automatiquement, en se servant de tests statistiques, la liste des descripteurs donnant la plus grande satisfaction à l'utilisateur (76, 118, 119, 149).

G) Evaluation des Systèmes Documentaires

L'efficacité d'un système documentaire se juge selon ses capacités de satisfaire les utilisateurs. Cleverson distingue six critères qui influencent la satisfaction des utilisateurs :

- la richesse du fond documentaire ;
- le temps de réponse moyen ;
- la présentation des résultats ;
- l'effort réclamé à l'utilisateur pour la formulation de sa question;
- le taux de rappel du système ;
- le taux de précision du système.

Les deux derniers critères sont les plus importants mais aussi les plus difficiles à évaluer. Le taux de rappel (recall) est la proportion de documents pertinents retrouvés lors d'une recherche. Le taux de précision est la proportion de documents retrouvés qui sont pertinents. Le taux de précision varie en fonction inverse du taux de rappel comme montre la courbe des performances d'un système documentaire.



Courbe rappel - Précision

Dans un système idéal, tous les documents pertinents seraient retrouvés et tous les documents non pertinents seraient rejetés ; la valeur rappel / précision serait égal à I. Malheureusement, ceci n'est pas le cas, l'extraction d'un grand nombre de documents pertinents s'accompagne de l'extraction d'un grand nombre de documents non pertinents.

Pour une question "Q" posée au système, on peut partager la collection documentaire comportant "N" documents en quatre sous-ensembles contenant respectivement a, b, c, et d documents comme dans le tableau suivant :

Tableau III : Réponses à une question Q

Documents	Pertinents	Non Pertinents	Total
Retrouvés	a	b	a + b
Non retrouvés	c	d	c + d
Total	a + c	b + d	a + b + c + d = N

Ce tableau permet de définir certains paramètres de mesure pour évaluer les performances d'un système documentaire en fonction des besoins

de ses utilisateurs.

. taux de rappel = $\frac{\text{nombre de documents pertinents retrouvés}}{\text{nombre de documents pertinents dans la collection}}$

$$R = \frac{a}{a + c}$$

. taux de précision = $\frac{\text{nombre de documents pertinents retrouvés}}{\text{nombre de documents retrouvés}}$

$$P = \frac{a}{a + b}$$

On peut également décrire un taux de bruit et un taux de silence. Par "bruit" on entend le nombre de documents non pertinents retrouvés par le système. Le "silence" est le nombre de documents pertinents qui ne sont pas retrouvés lors d'une recherche. Son évaluation est difficile car impossible à détecter.

. taux de bruit = $\frac{\text{nombre de documents non pertinents retrouvés}}{\text{nombre de documents retrouvés}}$

$$B = \frac{b}{a + b}$$

Deux autres paramètres sont utilisés :

. taux de chûte (Fallout) = $\frac{\text{nombre de documents non pertinents retrouvés}}{\text{nombre de documents non pertinents dans la collection}}$

$$F = \frac{b}{b + d}$$

. mesure de Mérite M = (a + d) - (b + c)

Tous ces paramètres sont liés entre eux. Les questions posées au système doivent être choisies pour faire intervenir tout le domaine d'activité et les diverses possibilités du système. Ainsi pour moduler les paramètres calculés pour les différentes questions on fait intervenir le degré de généralité de chaque question par rapport à la collection documentaire "N".

. facteurs de généralité : $\frac{\text{nombre de documents pertinents de la collection}}{\text{nombre total de documents de la collection}}$

$$G = \frac{a + c}{N}$$

En Janvier 1966, deux années après le début des opérations, un vaste projet d'évaluation a été commencé par le Medical Literature Analysis and Retrieval System (MEDLARS) de la National Library of Medicine (NLM) aux Etats-Unis. Le projet a pris deux ans pour se réaliser, prenant fin en Décembre 1967.

Vingt organisations médicales majeures ont participé au projet en fournissant 300 demandes de recherche.

Les buts ont été :

- d'étudier les besoins des utilisateurs du système ;
- de mesurer objectivement les performances du système en rapport avec ces besoins ;
- de découvrir des facteurs contribuant aux échecs du système ;
- de déterminer les moyens d'augmenter l'efficacité et l'économie du système.

On a fait appel aux six critères de Cleverdon pour l'étude des besoins des utilisateurs. On a établi un formulaire destiné aux demandeurs

.r

pour déterminer pourquoi certains articles sont jugés pertinents et d'autres articles non pertinents.

Les résultats ont montré que MEDLARS opère à un taux de rappel de 57,7 %, et à un taux de précision de 50,4 %. Ceci veut dire que le système retrouve un peu moins de 60 % des documents pertinents qu'il possède et que 50 % des articles retrouvés sont jugés pertinents par les utilisateurs.

Ces chiffres sont moins importants que les raisons des échecs du système. L'analyse de chaque échec a montré que les causes principales sont :

- le langage d'indexation (manque de termes spécifiques ou fausse coordination de termes) ;
- la stratégie de recherche (formulation des questions trop exhaustives ou trop spécifiques) ;
- l'indexation des documents (indexation trop ou trop peu exhaustive, omission de termes importants) ;
- le manque d'interaction utilisateur - système pendant la recherche.

Les résultats obtenus avec MEDLARS, où l'indexation n'est pas automatisée, ont été comparés avec les résultats du système SMART. Cette comparaison a révélé que le système SMART, complètement automatique, donne à peu près les mêmes valeurs pour le taux de rappel et le taux de précision que MEDLARS.

Dans l'étude de l'évaluation d'un système documentaire et dans la comparaison de deux systèmes il y a encore deux facteurs qu'il faut prendre en considération :

- l'efficacité "e" qui mesure la capacité du système à retrouver des documents pertinents vis-à-vis des questions posées par les utilisateurs ;

- le rendement "r" qui mesure le coût de fonctionnement du système en fonction du temps et de l'activité.

On voit que le coût des différentes opérations est fonction de la qualité des services demandés.

On peut calculer une fonction de performance "P" liant les deux facteurs efficacité et rendement :

$$P = \frac{e + r}{2}$$

Ce calcul permet de comparer, d'un point de vue global, les performances de deux systèmes différents.
(Références : 74, 75, 76, 77, 121, 122, 124, 142, 143, 155).

H) Le Coût de la Documentation Automatique

Il est très difficile de préciser le coût de la documentation automatique. Plusieurs facteurs interviennent dans l'appréciation de ce coût. Il y a d'abord le matériel informatique : la location de l'ordinateur et de ses périphériques (lecteur de cartes, dérouleur de bandes, unités de disques, imprimante, terminaux de visualisation).

Ensuite vient la gestion des documents et l'investissement intellectuel tel que la construction de dictionnaires, l'élaboration de grammaires, l'établissement de programmes. Il y a aussi les salaires des analystes à prendre en considération.

Même sans donner de chiffres il est facile d'imaginer l'ampleur du prix d'un tel système. Mais combien coûte une recherche documentaire

traditionnelle ? On estime que les savants, les ingénieurs et les techniciens passent environ 15 % de leur temps à des recherches documentaires, ce qui représente un coût d'environ 200 millions de dollars par an.

Un autre aspect du problème est le coût de l'analyse documentaire faite par les méthodes manuelles ou par les méthodes mécaniques. Si l'indexation automatique est techniquement possible, elle n'est pas encore économiquement compétitive avec l'indexation humaine. Pratiquement un document ne peut-être présenté à l'ordinateur que sous forme de cartes perforées ou de rubans papier perforé. Or la préparation de ces cartes ou de ces rubans est très coûteuse. La solution de ce problème attend soit le jour où tous les documents seront produits à la fois sous forme imprimée et sous forme directement lisible par la machine, soit le jour où les procédés de reconnaissance automatique d'un texte imprimé seront généralisés et universellement applicables. (14, 76, 149).

I) Les Systèmes Questions-Réponses Automatiques

Jusqu'ici on a considéré le cas où un utilisateur fait une demande en s'attendant à avoir comme réponse une liste de documents. L'indexation des textes était faite dans ce but. Cependant il existe des systèmes capables de fournir non pas des références mais des réponses directes aux questions posées. Il s'agit de systèmes question-réponse ou systèmes de recherche de données. Plus loin la banque de données en enzymologie sera longuement décrite, ici un bref exemple d'un système capable de répondre aux questions est exposé (102, 152).

Le Système ACORN (Automated Coder of the Report Narrative), développé par le Roswell Park Memorial Institute de Buffalo, pour traiter les comptes rendus opératoires, utilise un langage naturel. La base du système est une nouvelle unité d'information appelée "kernel" (noyau). Le kernel possède cinq propriétés désirables pour toute unité d'information :

1 - l'unité est commune à toutes les langues ;

2 - l'unité possède une définition opérationnelle formelle (l'objectivité est assurée par l'emploi d'un ordinateur qui extrait automatiquement les kernels des rapports) ;

3 - l'unité est pratique (l'analyse des rapports, la recherche rétrospective, et même la traduction automatique des kernels d'une langue à une autre sont possibles) ;

4 - l'information mesurée par les unités correspond étroitement au sens ordinaire du mot information (la longueur d'un message n'est pas confondue avec son contenu ; deux phrases structurallement distinctes contenant la même information donnent des kernels identiques) ;

5 - l'analyse en kernels d'un texte révèle les mécanismes grammaticaux avec lesquels l'information a été codée en langage naturel.

L'unité linguistique d'information - le kernel - est représentée par l'équation F (X) = Y où F est une fonction qui place un objet ou une action X dans un ensemble d'attributs Y décrivant X. Par exemple, si OP est la fonction qui signifie une opération par le chirurgien, OP (TUMEUR) = ENLEVEE est le kernel d'information contenu dans chacune des phrases suivantes :

1 - la tumeur a été enlevée ;

2 - la tumeur a été enlevée par le chirurgien ;

3 - le chirurgien a enlevée la tumeur.

Les kernels sont extraits automatiquement du texte. Chaque phrase passe par trois étapes dans l'analyse : consultation du dictionnaire, analyse syntaxique, et analyse transformationnelle, selon la méthode de Zellig Harris (string analysis).

Tous les kernels extraits sont stockés dans un "kernel catalogue", puis triés successivement par valeur X, F, et Y. Si une question est posée, telle que "y a-t-il des rapports mentionnant un examen pelvien ? "ou" dans quelle position a été placé le patient ? ", il est possible d'y répondre par le décodage de la question en kernels et la comparaison automatique de ces kernels avec ceux stockés dans le catalogue (121, 134).

J) Conclusion

Il est paradoxal que l'abondance de nos connaissances menace d'augmenter notre ignorance ; à mesure que nos informations accroissent il devient de plus en plus difficile d'extraire les faits importants. Dans la discussion précédente, j'ai essayé de démontrer qu'il existe des techniques automatiques destinées à aider cette recherche d'information pour l'accélérer et la rendre plus précise et moins pénible (4, 104, 137).

Pourtant les machines présentent des inconvénients. Le premier est leur manque de souplesse ; elles sont moins souples que le jugement humain quand il s'agit d'une indexation ou d'une recherche de document ; elles permettent moins facilement des modifications. Les ordinateurs nécessitent un capital initial important et leur entretien est coûteux. Ils peuvent tomber en panne et arrêter le fonctionnement de tout le système. Ils ne peuvent être exploités que par des opérateurs spécialement formés (150). Si la recherche documentaire automatique est devenue courante, l'analyse automatique est toujours un problème difficile.

A côté de cette vue un peu pessimiste, Salton a démontré, avec son système SMART, que les techniques documentaires automatiques, y compris l'analyse automatique, valent au moins les techniques manuelles. Pour Gardin "l'automatisation de l'analyse documentaire est une entreprise justifiée, du point de vue économique, sans que l'on ait à transiger sur la qualité finale du produit ; et tout porte à croire que la balance ne cessera de pencher

davantage en faveur de cette option, à mesure que le coût des machines baissera par rapport à celui du travail humain" (14).

Un problème reste entier du fait des difficultés sémantiques c'est le mauvais rendement d'une analyse documentaire quel que soit le procédé.

DEUXIEME PARTIE

Systèmes de Documentation Automatique Opérationnels et Spécifiques de la Médecine.

A) Introduction

Cette partie de la thèse concerne l'effort mondial actuel dans la documentation médicale automatique. Il n'est pas question de décrire tous les systèmes documentaires qui existent dans le monde. Après une étude de deux réalisations de grande envergure quelques exemples de travaux nationaux plus limités seront passés en revue, pour terminer avec la description d'un projet de l'O. M. S.

B) MEDLARS

MEDLARS (Medical Literature Analysis and Retrieval System), créé par la National Library of Medicine (USA), est devenu opérationnel en Janvier 1964 après trois années de développement. Son objectif principal est de fournir des références biomédicales aux chercheurs, aux cliniciens et aux autres professionnels de santé. Ceci est réalisé par :

- la préparation de citations pour la publication de l'Index Medicus et du Current Catalog ;
- la préparation des "Recurring Bibliographies", bibliographies périodiques contenant des références relatives à des domaines spécialisés ;
- la préparation des bibliographies rétrospectives (demand searches).

Trois principes classiques de la documentation et de la science de l'information ont gouverné la planification et l'opération du MEDLARS :

- l'utilisation d'un vocabulaire contrôlé pour l'indexation et le repérage dans le fichier ;
- l'utilisation d'analystes professionnels ;
- l'organisation séquentielle du fichier sur bande magnétique.

Le système est constitué de trois parties :

- un sous-système d'entrée ;
- un sous-système de recherche ;
- un sous-système de publications.

Un ordinateur Honeywell-200 est utilisé pour les opérations lentes d'entrée/sortie, tandis que le traitement des informations est réservé à l'ordinateur Honeywell-800. Les publications sont préparées pour l'imprimerie par le "Photon 900 computer Phototypesetter" appelé GRACE (Graphic Arts Composing Equipement).

Le sous-système d'entrée comprend la sélection des articles, l'indexation, la conversion en un format lisible par la machine, et l'entrée en ordination pour stockage sur bande magnétique.

La National Library of Medicine reçoit 18 000 publications de tous genres. De ceux-ci 2 300 périodiques biomédicales sont extraites pour l'indexation. En Juillet 1970 les bandes magnétiques de MEDLARS contenaient 1 254 200 citations environ. Chaque année approximativement 200 000 citations, dont la moitié ne sont pas en anglais, sont ajoutées. La sélection des journaux est faite avec le concours d'un groupe de médecins renommés, d'éditeurs médicaux et de bibliothécaires spécialisés dans le domaine médical.

Tout le système est basé sur un vocabulaire contrôlé appelé Medical Subject Headings (Me SH). Il s'agit d'une liste ou thésaurus de termes techniques (descripteurs) que les analystes utilisent pour l'indexation des articles, les recherches rétrospectives, et pour cataloguer des livres. Me SH est publié en Janvier de chaque année avec Index Medicus. Il est divisé en deux parties :

- la liste alphabétique contient les descripteurs et les références croisées arrangées en ordre alphabétique ;
- la liste des catégories affiche les descripteurs par grandes catégories montrant des rapports génériques.

En plus de ces descripteurs principaux, il existe un certain nombre de descripteurs géographiques, des descripteurs provisoires, et des descripteurs secondaires (subheadings). Dans l'exemple "KIDNEY NEOPLASMS-Surgery", néoplasme rénal est le descripteur principal (main subject heading) et chirurgie le subheading.

Ce thésaurus n'est pas statique, d'une part parce que Me SH est mis à jour périodiquement pour tenir compte des nouveaux concepts médicaux et des modifications de terminologie, d'autre part parce que le système de termes provisoires permet de tester un nouveau descripteur avant son inclusion définitive dans le thésaurus.

Certains journaux importants sont analysés "en profondeur", c'est-à-dire qu'on se sert d'une dizaine de descripteurs en moyenne pour décrire le contenu d'un article. Pour les autres journaux, seulement quatre descripteurs sont utilisés. Lancaster a montré que ce type d'indexation était une des causes du mauvais rendement du système. Il conseille l'utilisation d'autant de descripteurs qu'il est nécessaire pour exprimer le contenu d'un article.

Certaines règles gouvernent l'indexation. Il faut toujours utiliser le descripteur le plus spécifique, disponible. L'analyste doit surtout chercher dans l'article des éléments tels que : découvertes ou données originales, résultats négatifs significatifs, structures anatomiques, organismes, maladies, produits chimiques ou médicaments, techniques analytiques, méthodes thérapeutiques, aspects épidémiologiques, etc... De plus, il doit signaler : l'âge des malades étudiés, grossesse, expériences humaines ou animales, recherches cliniques, études de dossiers, facteurs géographiques.

Tous ces renseignements extraits de l'article y compris le titre, le nom de l'auteur, le titre de la revue, son numéro de volume, la pagination, la date, la langue, constituent l'unité d'information qui sera introduit en ordinateur et enregistré sur bande magnétique.

Le sous-système de recherche s'occupe de la comparaison des demandes de bibliographies avec les citations dans le "Compressed Citation File" (fichier principal). L'analyste est chargé de la formulation définitive de la question ; ensuite l'ordinateur compare la question avec les citations dans le fichier et imprime la bibliographie résultante.

Ce système est basé sur la logique de Boole avec ses opérateurs "ou", "et", "sauf", et sur un système spécial dans lequel les descripteurs sont ordonnés hiérarchiquement (Tree structures).

Il ne faut pas demander n'importe quoi au système. Une mauvaise question donne une mauvaise réponse. L'Index Medicus convient mieux à une demande telle que "tous les articles concernant les infections dues au virus Coxsackie". Mais une demande comportant la coordination de plusieurs concepts ; telle que "les effets des radiations sur les virus", est considérée valable pour une recherche MEDLARS. Il ne faut pas demander non plus : des recherches dans tout le fichier, des recherches de noms d'auteurs, des vérifications de citations, des données, des traductions.

Voici un exemple de formulaire utilisé par l'INSERM pour une demande de recherche bibliographique MEDLARS.

Formulaire pour questions documentaires posées au service recherches rétrospectives ;

1 - Nom du demandeur :
adresse :
grade ou fonction :
travail en cours pour lequel la question est posée :

II - Si vous avez déjà entrepris une recherche bibliographique sur le présent thème, énumérez les sources que vous avez utilisées :
Nombre de références connues de vous :
Citez 5 références caractéristiques.

III - Enoncé précis de la question par le demandeur :
Points qui vous intéressent particulièrement :
Objectifs de la recherche :

IV - Limitations souhaitées (langue, date des documents, types de sources, etc...) :

V - Ordre préféré pour les résultats : alphabétique d'auteur, alphabétique de titres de revues chronologiques par langue :
Urgence (date limite de réponse) :

VI - Inscrivez la liste des mots-clés :

La réponse fournie par l'ordinateur comprend :

- une liste de références bibliographiques ;
- pour chaque références les mots-clés qui décrivent le contenu de l'article.

Grâce à ces mots-clés attachés à chaque citation, le demandeur reçoit une indication du contenu de l'article cité.

Comment se déroule la préparation, par l'analyste, d'une question destinée à être soumise à l'ordinateur pour une recherche rétrospective ? La question suivante, par exemple, est reçue par un centre MEDLARS. "Nous travaillons sur les rapports entre les désordres du métabolisme protidique et les maladies du système nerveux. Toutes les références récentes nous intéressent, surtout celles en rapport avec les aminoaciduries et particulièrement la phénylcétonurie". L'analyste prépare sa stratégie ainsi : dans Me SH, "maladies du système nerveux" est une catégorie comportant 188 termes ; "désordres du métabolisme protidique" existe comme un terme, avec une liste de maladies spécifiques (ex : ochronose) ; on demande des rapports surtout avec les aminoaciduries et le phénylcétonurie.

Il arrive à dresser la liste des descripteurs appropriés :

M1 (maladies du système nerveux)
M2 désordres du métabolisme protidique
M3 albinisme
M4 albuminurie
M5 aminoacidurie
M6 alcaptonurie
M7 cystinurie
M8 maladie du sirop d'érable
M9 phénylcétonurie
M10 amyloidose
M11 cystinose
M12 ochronose
M13 (somme M2-M12)
M14 (somme M4-M9)

Sa stratégie définitive s'exprime symboliquement ainsi : (M1) * M13 qui veut dire : je veux toutes les citations dans la catégorie "maladies du système nerveux" en association avec chacun des descripteurs de M13 ;

M14 je veux toutes les références sur albuminurie, aminoacidurie, alcaptonurie, cystinurie, maladie du sirop d'érable, phénylcétonurie.

M9 toutes les références sur la phénylcétonurie.

Les questions sont introduites sous forme de cartes perforées, et groupées (une centaine) en vue d'un traitement simultané (Batch Processing). L'ordinateur compare la formulation de la recherche et les mots-clés décrivant les articles contenus dans le fichier. Les numéros correspondant aux références retenues sont alors triés par question, puis décodés. Ces références peuvent être organisées en plusieurs formats et imprimées sur papier continu ou sur cartes 75 X 125 mm.

Le sous-système de publication comprend des programmes et des procédures pour la préparation d'index périodiques de la littérature biomédicale courante et d'autres publications bibliographiques du MEDLARS. Parmi ces publications, signalons : Index Medicus, Medical Subject Headings, NLM Current Catalog, Cerebrovascular Bibliography, Index of Rheumatology, Current Bibliography of Epidemiology. Le Current Catalog est la liste des publications cataloguées par la National Library of Medicine. Les bibliographies périodiques sont établies régulièrement à partir des données contenues dans la mémoire de l'ordinateur et distribuées aux abonnés de ce service.

La décentralisation du MEDLARS, commencée depuis 1965, est maintenant internationale. Il existe onze centres MEDLARS régionaux aux Etats-Unis et huit en dehors du pays, Le réseau international de centres MEDLARS comprend les pays suivants : Australie, Canada, Grande-Bretagne, France, Allemagne, Japon, Suède et Suisse. Cette collaboration nécessite l'acceptation d'une discipline commune pour l'indexation, et l'utilisation du même dictionnaire.

Après les accords conclus en 1969 avec la National Library of Medicine, l'Institut National de la Santé et de la Recherche Médicale (INSERM) reçoit une copie des bandes magnétiques MEDLARS. En échange l'INSERM fournit l'analyse et l'indexation (suivant les critères définis par le NLM) de 140 revues biomédicales de langue française. Ainsi l'INSERM peut offrir aux enseignants et aux chercheurs français deux types de services :

- des recherches bibliographiques rétrospectives sur la totalité du fichier (articles parus depuis 1964) ou une partie plus récente du fichier ;
- une diffusion sur "profil" sous forme d'un abonnement annuel qui procurera chaque mois une liste de références pour un sujet donné.

La question de l'évaluation de MEDLARS a été étudiée dans la première partie de cette thèse. D'autres critiques du système ont été avancées. MEDLARS ne produit pas des résumés et ne repère que la littérature publiée dans les périodiques. Une fois muni de la référence l'utilisateur doit faire face au problème de la recherche de l'article original. Les intéressés ne peuvent pas questionner directement le fichier. Ils doivent s'appuyer sur des intermédiaires, sur des spécialistes qui transcrivent et formulent leurs demandes dans le langage de l'ordinateur. L'efficacité de repérage est déterminée par la spécificité des Me SH et l'habilité des analystes. Un terme spécifique n'existe pas toujours et dans certaines disciplines (orthopédie par exemple) les Me SH ne sont pas adéquats. On manque du personnel spécialisé dans l'indexage, tandis que la littérature biomédicale ne cesse de s'accroître. La logique de Boole ne couvre pas toutes les situations.

Ces critiques et les résultats de l'évaluation de Lancaster ont abouti à l'installation de MEDLARS II. Ce projet est actuellement en cours de réalisation aux Etats-Unis, à la National Library of Medicine ; son exploitation sera possible sur le mode conversationnel.
(Références : 3, 5, 9, 10, 18, 21, 28, 29, 46, 50, 55, 57, 61, 65, 67, 74, 86, 98, 99, 100, 109, 111, 114, 115, 116, 131, 133, 144, 145, 146, 151).

C) Excerpta Medica

Fondée en 1946, la Fondation Excerpta Medica publie à Amsterdam 34 séries de résumés en anglais, chacune représentant une spécialité particulière de la médecine (Médecine Interne, Ophthalmologie, Pédiatrie, Radiologie) ou un domaine interdisciplinaire (Réhabilitation, Médecine Nucléaire, Gérontologie, Gériatrie). Ainsi, sont examinées chaque année 3 000 revues médicales, soit 20 000 publications annuelles ou 200 000 articles. L'analyse des articles est faite par des médecins spécialistes du sujet pour assurer une haute qualité de l'indexation. Ceci représente une collaboration de 4 000 médecins dans le monde entier, sous la supervision d'un conseil scientifique siégeant à Amsterdam.

L'Université de Leyde, l'Académie Royale des sciences des Pays-Bas et la Fondation Excerpta Medica coopèrent à la constitution d'une bibliothèque où sont archivés les microfilms de chaque publication avant le travail d'abstraction et d'indexation. En plus de la compilation et la production de la série d'Excerpta Medica, le système offre un service de signalisation pour la littérature courante et de recherche bibliographique rétrospective . Les bandes magnétiques de ces deux services peuvent être achetées. Les programmes sont prévus pour fonctionner par lots ou en interrogation directe, en temps réel. Une centaine de périphériques "on-line" est reliée au système central permettant des consultations très rapides.

Le vocabulaire est contrôlé ; 500 000 termes (synonymes et variations dans différentes langues) se refèrent à une collection de 40 000 termes principaux, couvrant l'ensemble des notions bio-médicales. Le repérage dans ce thésaurus est automatique. En plus des 40 000 termes principaux, le système utilise un certain nombre de termes non-contrôlés pour décrire le caractère de la recherche, par exemple : anatomique, démographique, etc...

Finalement, des centres régionaux sont établis dans plusieurs pays pour permettre le traitement local par ordinateur des informations destinées à entrer dans le système. Ce service est disponible aux médecins-chercheurs, aux institutions médicales et aux centres d'information (1, 35, 44, 66, 70, 73).

D) Le Système du Cox Coronary Heart Institute.

Ce Système de documentation automatique, développé par le Cox Coronary Heart Institute aux Etats-Unis, peut répondre aux questions spécifiques et fournir des bibliographies, des abstracts, ou des articles originaux (publiés depuis 1960) sur les maladies cardiaques coronariennes. Le résumé et le document original peuvent être extraits presque simultanément, "on-line" et en temps réel, et dans une forme transmissible à un utilisateur même très éloigné.

Les résumés, d'environ 400 mots chacun, sont préparés par des spécialistes ; 110 périodiques sont analysés et 175 résumés en moyenne sont produits par mois. Les mots-clés extraits du titre et du résumé de l'article sont comparés visuellement avec un lexique de mots-clés pour permettre seulement l'introduction de mots pertinents. Ensuite les mots-clés sont rassemblés par la machine dans plusieurs groupes et couplés avec d'autres mots-clés et phrases contextuelles. Les mots "a", "and","of", et "the", sans valeur pour les recherches rétrospectives, sont éliminés automatiquement du résumé et tous les mots restants sont considérés comme mots-clés.

Une recherche sur une question spécifique s'opère de la façon suivante. Voici par exemple, une question provenant d'un laboratoire travaillant avec des chiens : Pourriez-vous nous fournir des renseignements concernant la température des ventricules du coeur (heart) chez le chien (dog) ? Cette question (en anglais) est posée à l'ordinateur ainsi : CANINE, DOG, DOGS ; TEMPERATURE, HEAT, COLD ; VENTRICLE, HEART. Les vir-

gules entre les mots-clés impliquent "ou" et le point virgule implique "et". Un point signale la fin de la question. La recherche des documents dans le fichier se fait en trois étapes : d'abord une recherche sur les mots "canin", "chien et" "chiens" ; ensuite en se limitant à ce groupe de documents, les mots "température", "chaleur", et "froid" sont cherchés ; la dernière recherche, sur les documents restants, est faite pour les mots "ventricule" et "coeur".

Si les résumés qui en résultent ne sont pas suffisants, un accès immédiat aux documents originaux est possible grâce à une technique photomicrographique. 1 200 pages des documents originaux sont inscrites sur une seule carte de 46 cm. Une visionneuse spéciale permet la lecture de l'article après un repérage utilisant des coordonnées numériques x-y.

Ce système de documentation automatique appliqué à une littérature très spécialisée pourrait s'adapter à d'autres branches de la médecine (147).

E) SABIR

Le Système Automatique de Bibliographie, d'Information et de Recherche en Carcinologie (SABIR), opérationnel depuis 1968, a été créé par le service de Documentation de l'Institut Gustave Roussy pour traiter les informations en cancérologie clinique et expérimentale. 20 000 documents, paraissant dans 1 000 revues, sont analysés par an. Le thésaurus prend en compte des relations de synonymie, de hiérarchie et de voisinage relatives à chaque concept. Pour permettre la coopération internationale, le système a trois langues de travail (français, anglais et allemand), avec passage automatique de l'une à l'autre.

Le dictionnaire est périodiquement remis à jour. Un programme spécial permet de rattacher automatiquement les mots nouveaux acceptés par

le système avec les autres termes du thésaurus. Un document est caractérisé par une dizaine de descripteurs principaux et une dizaine de descripteurs accessoires (langue de parution). Ce travail d'analyse est assuré par des spécialistes en cancérologie.

Le système permet : la publication d'une revue mensuelle bibliographique "Cancer", éditée à partir des listes fournies par l'ordinateur ; une exploitation sur profil d'intérêt ; et des recherches rétrospectives sur question particulière.

L'institut bénéficie d'une large coopération internationale, surtout avec le Centre Allemand de Recherche sur le Cancer d'Heidelberg, l'Institut d'Oncologie de Varsovie et le Centre International de Recherche sur le Cancer de Lyon. Ainsi s'est formé un réseau de centres anticancéreux dans plusieurs pays coopérant à la constitution d'un pool d'analyses de la littérature (73, 117, 154)

F) DIMDI

L'Institut Ouest-Allemand pour la Documentation Médicale et l'Information (DIMDI) fonctionne à Bonn depuis 1969. Ce vaste complexe d'ordinateurs utilisé pour la documentation médicale automatique, se sert du système MEDLARS et des méthodes de classification de l'Index Medicus. En plus DIMDI prépare et stocke des résumés sur les articles les plus importants en médecine parus depuis 1964 (34).

G) FAIR

Semblable à MEDLARS, Fast Access Information Retrieval (FAIR) est un projet du National Institute of the British Medical Research Council. Le domaine du FAIR comprend des journaux et des livres dans toutes les spécialités médicales et chirurgicales, la santé publique et l'hygiène, les sciences biologiques alliées à la médecine et le biomedical engineering. Son

but est, avec le concours d'un ordinateur, le stockage et l'extraction faciles et rapides des informations sous forme de résumés et éventuellement sous ses formes originales : articles et documents entiers (110).

H) L'O. M. S. et la Documentation Automatique

L'organisation mondiale de la santé à un système de recherche rétrospective des informations qui fournit des résumés de tous les paragraphes des rapports concernant le sujet de la demande. Chaque document reçoit un numéro à quatre chiffres. Les deux chiffres de gauche correspondent aux grandes divisions de l'oeuvre de l'O. M. S. Donc Maladies Infectieuses, Services Publics de Santé etc... sont signalés par 1 000, 2 000, 3 000 ..... ; et Tuberculose (sous Maladies Infectieuses) et Reproduction Humaine (sous Biologie) par 100, 200, ... L'éditeur analyse chaque document et le classe par ces deux chiffres, puis chaque paragraphe reçoit les deux chiffres de droite du numéro de code. Ensuite, il fait le résumé de tous les paragraphes comportant les mêmes chiffres de droite. Un tel résumé constitue l'unité d'information ou "module". Voici un exemple : on demande des renseignements sur les "recherches dans l'efficacité comparée des méthodes de communication utilisées pour l'enseignement de l'hygiène" ; pour y répondre il suffit d'entrer dans l'ordinateur une carte perforée avec le numéro 4555, c'est-à-dire, 4000 pour services publics de santé, 500 pour enseignement de l'hygiène, 50 pour recherche et 5 pour l'efficacité comparée des méthodes de communication (62).

I) Conclusion

Tous ces systèmes de documentation automatique au service de la médecine doivent augmenter le rendement des médecins. L'ordinateur doit libérer le cerveau de l'homme tout comme les machines énergétiques ont libéré ses muscles (26). Les machines informationnelles, qui n'oublient rien, qui effectuent ses tâches avec une rigueur et une rapidité inaccessibles à l'homme, qui exécutent plusieurs opérations simultanément, vont libérer le

médecin de servitudes pénibles, accroître l'efficacité de ses gestes diagnostiques et thérapeutiques, et rendre son cerveau à son véritable champ d'action : la réflexion morale et philosophique (16, 97).

TROISIEME PARTIE

Services Documentaires Automatiques Non Spécifiquement Médicaux.

A) Introduction

Le premier périodique scientifique, le "Journal des savants", est né en 1665. Deux cents ans plus tard, en 1876, on en comptait 4 390. Les communications scientifiques continuaient de proliférer, suivies à leur tour par les publications secondaires des résumés analytiques et des index avec : Chemisches Zentralblatt en 1830, Index Medicus en 1879, Chemical Abstracts en 1907. Entre 1907 et 1916, Chemical Abstracts ont produit 180 000 résumés analytiques. En 1967, cinquante ans plus tard, ils ont produit 310 000.
Il en est de même pour Biological abstracts qui, de 30 000 résumés en 1930, ont produit 100 000 en 1963. Aujourd'hui il y a environ 30 000 à 50 000 revues scientifiques publiées par an, dont 18 000 en médecine ; les revues des résumés analytiques et index sont estimées à 2 000 publications (70). Devant cette avalanche de documents l'intervention de l'ordinateur est devenue une urgence.

B) Biological Abstracts, BASIC Index, et Bio Research Index.

La publication de "Biological Abstracts" a débuté en 1926. Les résumés de la littérature en Biologie apparaissent deux fois par mois et traitent, par exemple, la bactériologie, la biochimie. Les articles y sont répertoriés par sujet ; dans le "BASIC index" ils le sont par titre permuté. Chaque année, sont publiés des index cumulatifs d'auteurs et de sujets; depuis 1965 un Bio Research Index de 7 500 articles concernant la recherche biologique mondiale est publié tous les mois (I).

Le BASIC Index est la clé de la recherche mondiale en biologie. Il s'agit de titres d'articles permutés selon leurs mots-clés comme dans un KWIC index. Les titres sont inscrits sur cartes perforées et données à l'ordinateur pour le travail de rangement et d'impression du résultat. L'utilisateur peut donc, à son bureau même, choisir les mots-clés selon ses propres spécifications (95). Avec un système comme Tabledex on peut faire mieux encore ; l'ordinateur produit cette bibliographie personnalisée qui est ensuite envoyée à l'utilisateur (158).

C) Chemical Abstracts, Chemical Titles, et Chemical Abstracts Condensates.

Chaque semaine depuis 1907 l'American Chemical Society publie des résumés de la littérature mondiale en chimie, groupés en 80 catégories : Biochimie, Chimie macromoléculaire, Chimie organique, etc... Les résumés sont répertoriés par mot-clé, formule chimique, numéro de brevet, et auteur (I). Environ 270 000 résumés sont publiés par an et couvrent plus de 12 000 revues (provenant de vingt-six pays différents) ainsi que des livres, des comptes-rendus de conférences, des rapports des agences gouvernementales (149).

L'ordinateur est entré au Chemical Abstracts Service en 1960 avec la publication, deux fois par mois, d'un KWIC index appelé "Chemical Titles". Ainsi l'intéressé est averti des titres d'articles paraissant dans 600 journaux de chimie pure et appliquée. Chaque semaine depuis 1968, des Chemical Abstracts Condensates sont disponibles sur bande magnétique ; ils contiennent les titres, les noms des auteurs, les références bibliographiques citées et les éléments du KWIC pour chaque article, rapport et brevet analysés par Chemical Abstracts (II). Les bandes magnétiques, directement lisibles par la machine, permettent à chaque utilisateur d'écrire ses propres programmes pour effectuer des recherches documentaires automatiques (149).

D) Science Information Exchange (SIE)

L'Institut Smithsonian, Washington, D. C., par l'intermédiaire du S. I. E., offre un service d'informations sur les activités de recherches en cours, non publiées, dans les sciences pures et appliquées, y compris la médecine et la biologie. Les "entrées", provenant de sources gouvernementales, privées et universitaires, sont enregistrées et rappelées à l'aide des ordinateurs sur demande. On peut obtenir des informations sur l'annonce du projet de recherche, son titre, un résumé de 200 mots sur les détails techniques, les noms et les adresses des principaux chercheurs, le nom de l'agence de subvention, et le niveau de la subvention (1, 70).

E) Automatic Subject Citation Alert (ASCA)

Depuis 1965, l'Institute for Scientific Information (I. S. I.) fournit un service de "dissemination sélective de l'information" appelé : ASCA. Le service compare automatiquement le "profil d'intérêt" de l'utilisateur avec les titres des articles publiés chaque semaine dans la littérature scientifique. Ainsi un chercheur pourrait recevoir régulièrement une bibliographie personnalisée sur les documents les plus récents publiés dans son domaine.

Ce système de signalisation hebdomadaire fonctionne par mots-clés non déterminés (KWIC), donc par un vocabulaire libre. Les mots significatifs du titre et sous titres des articles publiés dans plus de 2 400 périodiques scientifiques sont enregistrés sur bandes magnétiques avec les informations bibliographiques correspondantes. L'abonné doit faire connaître au service le sujet et les noms d'auteurs ou d'organisations qui l'intéressent, en même temps que quelques mots-clés et phrases décrivant son sujet, le titre d'un ou deux articles ou livres qu'il considère pertinents et, s'il a écrit sur le sujet lui-même, le titre, la date et le nom du journal qui a publié son article. L'ordinateur fait le reste. Le prix de l'abonnement est de 115 dollars par an (1, 70).

En plus de la rapidité et de l'étendue du service, le système offre l'avantage de "selectivity". On peut recevoir seulement : les articles écrits par certains auteurs, paraissant dans certaines revues, ou écrits par des hommes travaillant dans certaines organisations ; des articles contenant certains mots ou fragments de mots ; ou des articles qui citent un journal particulier, un auteur ou une organisation.

F) Science Citation Index (SCI)

Disponible depuis 1961, le SCI publié par l'Institute for Scientific Information est essentiellement un index des auteurs. Son but est de permettre aux chercheurs d'une même discipline de se reconnaître et de trouver les publications qui les intéressent (I). L'index est une liste de références arrangées par ordre alphabétique des noms d'auteurs, et suivies de tous les noms d'auteurs qui ont récemment cité ces références. Un index des sources donne le titre de ces articles citant (51, 90, 153). Il se construit ainsi un réseau de citations qui lie des documents par leurs bibliographies (136).

L'intérêt de cette "indexation par citation" et de ces "couplages bibliographiques" réside dans l'hypothèse que la bibliographie d'un article indique son sujet (120), qu'il existe un rapport entre l'article citant et l'article cité, et que ces éléments pourraient être utiles dans une recherche rétrospective de documents. Salton et Kessler (68, 69, 120) ont étudié cette question et pour eux l'hypothèse s'est révélée valable.

G) Current Contents, Life sciences. Permuterm Subject Index (PSI)

Pour être complet, notons que l'ISI publie chaque semaine depuis 1958 les Current Contents, Life sciences. Il s'agit de tables de matières, dans leur format original, de plus de 900 journaux. Un répertoire des noms et adresses des auteurs est inclus pour faciliter la demande de tirés à part ; ceux-ci peuvent également être fournis par l'ISI.

Depuis 1966, l'ISI publie le Permuterm Subject Index. Le PSI est un index des sujets de la littérature scientifique et technologique. L'index de 1968 concerne 300 000 articles publiés dans 1 711 journaux. Les mots significatifs des titres et des sous-titres des articles sont permutés pour former toutes les paires de termes possibles. Ainsi un "champ d'indexation" est établi qui permet à l'utilisateur de trouver des informations sur toutes les combinaisons de termes qu'il désire (I). Il suffit d'un seul mot pour faire une recherche de documents. L'index en langage naturel, est multidisciplinaire et international ; ses sources proviennent de 100 disciplines majeures et d'environ 50 pays.

L'ISI est aussi en mesure de faire des recherches rétrospectives de la littérature scientifique des 10 dernières années pour trouver tous les articles concernant un sujet donné.

H) Conclusion

Toute cette activité des publications secondaires démontre une duplication de l'effort rivalisant celle des documents primaires. Un même article peut-être analysé à la fois par les services de l'ISI, Biological Abstracts, Chemical Abstracts et Index Medicus. Si l'on considère que plus de la moitié des articles analysés par Chemical Abstracts (73 %) et Index Medicus (45 %) vient de sources en dehors des Etats-Unis et que d'autres pays s'occupent de cette même littérature, on voit que cette situation n'est certainement pas économique et dans une certaine mesure, ne répond pas non plus aux besoins des utilisateurs. La solution vraiment efficace à nos problèmes de communication scientifique et technique réside dans un effort coordonné international et multidisciplinaire. Plusieurs projets utilisant la technologie de l'informatique sont actuellement en plein développement pour résoudre ces problèmes : la décentralisation de Medlars, l'association de Chemical Abstracts et Chemisches Zentralblatt, et la coordination des procédés d'analyse d'Index Medicus et de Chemical Abstracts permettant ainsi la conversion automatique entre les deux systèmes d'indexation (11, 23, 84, 85, 101, 125).

QUATRIEME PARTIE

La Banque de Données

A) Introduction

Il faut distinguer les systèmes documentaires où les éléments traités sont des articles, des résumés ou des citations bibliographiques, et les systèmes qui traitent des informations numériques et non numériques destinées à former une banque de données (49, 76, 154).

"Les banques de données réduisent l'information en particules les plus fines possibles et fournissent non plus des références à des documents qu'il faut encore consulter, mais les informations élémentaires et les méthodes de traitement de ces informations qui permettent d'en tirer des conséquences ou de faire apparaître des théories nouvelles" (149, p. 120).

Le tableau suivant (Salton 121, p. 394) montre les principales différences entre les systèmes de recherche de références et les systèmes de recherche de données (traduction de Dr. Josette Zeraffa).

Tableau IV : Différences entre recherche de références et recherche de données

Recherche de références	Recherche de données
Les données de base sont des documents identifiés par des ensembles de concept sans liens logiques entre eux.	Les données de base sont des classes d'éléments pouvant être associés ou liés par des relations fonctionnelles.
Les données de base forment un système clos puisque toutes les réponses possibles sont déjà stockées.	Les données de base forment un système ouvert impliquant des procédures déductives susceptibles d'engendrer de nouvelles propositions à partir de données originelles.
Un seul type de question	Plusieurs types de questions
L'identification se fait par la simple recherche de correspondances entre les concepts des questions et des documents.	Les relations spécifiques entre les concepts de la question doivent d'abord être évaluées dans la base de données avant qu'une réponse ne soit fournie.
Un nombre restreint de relations entre les objets est spécifié : synonyme et/ou, relation de hiérarchie.	Sont spécifiés divers types de relations fonctionnelles entre les objets (par ex. : relations de classe, relations arithmétiques, etc...).
On néglige les associations entre les objets exprimées par des verbes et en général les termes exprimant une fonction.	Les relations entre objets doivent être spécifiées avant qu'une réponse ne puisse être élaborée.

B) Systèmes de Recherche de Données

Les systèmes de recherche automatique de données ou systèmes question-réponse nécessitent :

- une analyse en profondeur des documents ;
- un vocabulaire normalisé (thésaurus) ;
- une syntaxe permettant d'expliciter toutes les relations existant entre plusieurs notions dans un document donné;
- un traitement en ordinateur (157).

Ces systèmes, qui s'appliquent pour la plupart à des domaines nettement délimités, permettent de poser des questions de plusieurs types. Ils utilisent des procédures d'identification complexes où les données enregistrées et organisées en classes pourront être considérées premièrement comme identificateurs (descripteurs, caractéristiques), correspondant aux éléments donnés de la question, deuxièmement comme identifiables (éléments recherchés répondant à la question), multipliant ainsi les modes d'accès (157).

On peut définir une banque de données comme étant un système informatique permettant d'obtenir rapidement des réponses claires et exactes à tout type de question portant sur le contenu des documents analysés (48).

C) La Banque de Données en Enzymologie

Entre 1967 et 1969 le service de documentation de l'I N S E R M a constitué une banque de données en enzymologie. L'analyse en profondeur est faite par des spécialistes de la discipline. Le système est basé sur un ensemble de programmes dénommé S I N B A D (Système Informatique pour Banque de Données) mis au point sur U N I V A C 1 107 dans le cadre d'un contrat D G R S T visant à l'établissement d'une banque de données scientifiques, accessible en temps réel (48). Le programme des recherches a été

conçu de façon à être indépendant du contenu des données et pourrait être appliqué à tout autre domaine, scientifique ou non scientifique (12, 48).

L'analyse des articles fait appel à un code constitué de "grilles d'analyse". Ces grilles d'analyse comportent un ensemble de "colonnes" rattachées à un thème central (titre de grille) : nature de l'enzyme, réaction enzymatique, effet de la température sur l'activité enzymatique, enzymopathies, comportement immunologique des enzymes, etc... (157).

Les colonnes regroupent des "classes de variables" dont le caractère commun est donné par le titre de colonne (157).

Les variables correspondant au titre des colonnes peuvent être soit des "caractéristiques" (classes dont le nombre de termes est réduit et prédéterminé), soit des "individus" (classes dont le nombre de termes est illimité), soit des données numériques (48, 157).

Pour coder l'information contenue dans un document, l'analyste isole des unités d'information cohérente (séquences) qui seront reportées sur les grilles pertinentes. L'ensemble des grilles fonctionnera ainsi comme un questionnaire. Un contrôle de pertinence est effectué à la réception des analyses (157).

Le programme de chargement des données comprend quatre phrases :

1 - séparation des classes d'individus et des classes de caractéristiques dans les grilles ;

2 - établissement de la fréquence des caractéristiques dans les colonnes et constitution du "bloc lignes" en remplaçant les caractéristiques par leurs rangs de fréquence ;

3 - établissement de divers fichiers d'individus (R E F C O D, R E F C L A I R), avec les listes de lignes représentées pour chaque individu, et leur dénomination en clair ;

4 - établissement du fichier "C O D C L A I R" qui permet de décoder les titres des colonnes et les caractéristiques (157).

Les questions sont introduites sous formes chiffrées ; l'utilisateur doit recourir à un dictionnaire et observer quelques règles simples de grammaire. Les réponses reproduisent les titres des colonnes et des grilles (157).

Le programme d'interrogation écrit en langage d'assemblage pour U N I V A C 1 107 comprend quatre parties :

1 - diagnostic de la question pour vérifier si elle est correctement codée et pour la décomposer en ses éléments constitutifs (question élémentaire ou composée) ;

2 - décodage des questions et des réponses ;

3 - traitement des questions élémentaires ;

4 - traitement des questions composées (157).

Par comparaison avec le système américain T. D. M. S. (Time Shared Data Management System), les programmes SINBAD permettent "une très sensible économie d'écriture dans les fichiers et, par suite, une nette réduction des investissements en mémoire à accès direct" (12).

D) Conclusion

Les banques de données permettent d'espérer échapper à l'étouffement dû à la multiplication des références fournies par une documentation automatique (149). Rappelons qu'on attend la publication cette année de 300 000 articles dans le domaine biomédical, que le chercheur russe consacre déjà 22 % de son temps à la lecture, le chercheur américain 34 %, et

qu'on n'est jamais sûr d'avoir lu tout ce qui est paru dans un domaine particulier (154). Une question fondamentale se pose : comment obtenir toute l'information dont on a besoin mais juste celle qu'on est en mesure d'assimiler (149) ?

Lors d'un congrès sur l'Informatique en Médecine (au Canada en 1969) Bertrand Kempf de la Compagnie I B M, France, rapporte que la justification des banques de données tient au développement accéléré des connaissances médicales. Aujourd'hui il faut des mois ou des années pour que la découverte d'une équipe de recherche atteigne la masse des médecins. Demain, "il suffira de quelques instructions modifiées dans un programme, et, dans la minute, tous les médecins abonnés à l'ordinateur seront à même d'en profiter". Les praticiens, s'adressant quotidiennement aux centres d'informatique à propos de problèmes concrets, seront constamment tenus au fait de la plus récente actualité. Ainsi ils recevraient une information permanente rendant tout enseignement post-universitaire inutile (56, 108).

CONCLUSION

Plusieurs systèmes de documentation médicale automatique sont déjà opérationnels.

Le système MEDLARS américain est devenu international et se transforme actuellement en MEDLARS II. Il existe des centres de documentation automatique dont le rôle essentiel est de fournir des bibliographies à la demande ou sur profil d'intérêt.

Tous ces systèmes et services utilisent une méthode mécanique faisant appel aux ordinateurs pour les recherches de documents, et une méthode manuelle ou une méthode automatique rudimentaire (index KWIC) pour l'indexation. L'efficacité d'un système de documentation dépend, d'une part de la qualité du langage d'indexation et de l'analyse documentaire, d'autre part de l'organisation des fichiers et des facilités que l'ordinateur fournira pour leur gestion ; elle est évaluée par la pertinence et le coût des résultats obtenus.

Le problème le plus difficile à résoudre pour l'automatisation complète est l'analyse automatique des documents. Le projet SMART de Salton semble offrir les meilleurs espoirs dans cette voie et donne des résultats au moins égaux aux techniques manuelles.

Finalement les banques de données dépassent la simple recherche automatique de références bibliographiques et permettent de donner aux questions posées des réponses directes et précises, dans des délais courts.

Vu, le Président de Thèse :

Professeur H. PEQUIGNOT

Vu, le Doyen de la Faculté :

Professeur DELBARRE

Vu et Permis d'Imprimer :

Le Recteur de l'Académie de Paris

R. MALLET

I - BIBLIOGRAPHIE SELECTIVE

1 - ABENDSCHEIN W.F. -

Orthopaedics and the information explosion.
A problem in communication J. Bone and Joint Surg Vol.-52-A, n° 1, pp. 184-193. January 1970.

2 - ACKERMAN E., ROSEVEAR J.W., JUERGENS J.L. -

Storage and retrieval of biomedical data.
Paper 15 th. Ann. Conf. on Eng. in Med and Biol., Chicago November 1962.

3 - ADAMS S. -

MEDLARS : performance, problems, possibilities
Bull. Med. Libr. Ass. 53 : 139-151. April 1965.

4 - ADAMS S. -

The medical literature crisis -
past and present Med. Doc 5 : 93-95. October 1961.

5 - ADAMS S., TAINE S.I. -

Searching the medical literature. Information retrieval (MEDLARS) at the National Library of Medecine J A M A 188 : 251-254. April 20, 1964.

6 - AINSWORTH G.C. -

Storage and retrieval of biological information.
Nature 191 : 12-14, July 1961.

7 - ALOUCHE F., BELY N., CROS R.C., GARDIN J.C., LEVY F., PERRIAULT J. -

Economie Général d'une Chaîne Documentaire Mécanisée.
Gauthier-Villars, Paris 1967.

8 - ANDERSON J. -

The computer : medical vocabulary and information.
British Medical Bulletin Vol. 24. N° 3. pp. 194-197. September 1968.

9 - AUSTIN C.J. -

Data processing aspects of MEDLARS
Bull. Med. Libr. Ass. 52 : 159-163. January 1964.

10 - AUSTIN C.J., MEDLARS 1963-1967. -

National Library of Medicine, Bethesda. Md. 1968. 76 pages.
(Public Health Service Publication n° 1,823).

11 - BAKER D.B. -

Communication or chaos ?
Science vol. 169,PP 739-742. August 21. 1970.

12 - Banque de Données de l'I.N.S.E.R.M.

Rev. Informatique
Méd. n° 2. p. 177. juin 1970.

13 - BARR K.P. -

Estimates of the number of currently available scientific and technical periodicals.
J. Doc. Vol. 23. n° 2 June 1967.

14 - BELY N., BORILLO A., VIRBEL J., SIOT-DECAUVILLE N. -

Procédures d'Analyse Sémantique Appliquées à la Documentation Scientifique Gauthier-Villars.
Paris. 1970, 240 pages.

15 - BLAIN G. -

L'informatique et les archives médicales. L'Union Médicale du Canada tome 99. pp. 314-317.
février 1970.

16 - BLAIN G. -

Les robots parmi nous.
L'Union Médicale du Canada tome 26. pp. 1.546-1.550. décembre 1967.

17 - BORKO H., BERNICK M.D. -

Automatic document Classification.
J. ACM Vol. 10, n° 2. pp. 151-162. April 1963.

18 - BREHANT J., VERMEIL G. -

Où en est notre informatique médicale ?
La Presse Médicale 70. n° 19. pp. 895-898, 10 avril 1970.

19 - BROWNSON H.L. -

Evaluation of document searching systems and procedures
J. Doc. Vol. 21. n° 4. December 1965.

20 - BROWNSON H.L. -

Research on handling scientific information.
Science Vol. 132. nº 3 444. December 30. 1960.

21 - BRUNET P. -

Classification et Thésaurus.
La Presse Médicale 78. nº 46. p. 2.032. 31 octobre 1970.

22 - BRUNET P. -

Médecine et calculateurs.
La Presse Médicale. 76. nº 12. p. 525. 9 mars 1968.

23 - CARTER L.F. -

National document - handling systems in science and technology
Science vol. 154. nº 3 754. december 9. 1966.

24 - Ce n'est pas encore demain que les ordinateurs traduiront les articles de médecine.

La Presse Médicale 75. Nº 36. p. 1.794. 9 septembre 1967.

25 - CHESHIER R.G. -

Machine information search system.
Bull. Med. Libr. Ass. 50 : 481-486. July 1962.

26 - COUFFIGNAL L. -

La Cybernétique Presses Universitaires de France.
Paris 1963. 125 pages.

27 - CROS R.C., GARDIN J.C., LEVY F. -

L'automatisation des Researches Documentaires. Un Modèle Général : Le Syntol (2ed.) Gauthier-Villars.
Paris 1968. 260 pages.

28 - CUMMINGS M.M. -

A library-based computer system for indexing, storing and retrieving literature in the cardiovascular field. Ann. N.Y. Acad. Sci 126 : 919-25.
August 6. 1965.

29 - Cumulated Index Medicus J.A.M.A. 193 : 130-131, July 19, 1965.

30 - DAMERAU F.J. -

An experiment in automatic indexing Am. Doc. Vol. 16. nº 4.
October 1965.

31 - DELAVENAY E. -

La Machine à Traduire. Presses Universitaires de France. Paris 1963. 128 pages.

32 - DENNIS S.F. -

The construction of a thesaurus automatically from a sample of text. Symposium on Statistical Association Methods for Mechanized Documentation Natl. Bur. Std. (U.S.), Misc. Publ. 269. December 1965.

33 - DENNIS S.F. -

The design and testing of a fully automatic index-searching system for documents consisting of expository text in G. Schecter (ed.) Information Retrieval. A Critical View, Thompson Book Co., Washington. D.C. 1967.

34 - DIDI starts activities JAMA vol 210, n° 3 p 566, October 20, 1969.

35 - DIMITRIE D., NANU D. -

La presse médicale et le traitement moderne de l'information. La Revue du Praticien tome 19. n° 22. pp. X-XI. I septembre 1969.

36 - Documentation, indexing, and retrieval of scientific information U.S.

Senate Comm. on Govt. Operations, Senate Document. 113. 86 th. Congress, 2 nd. Session. U.S. Govt Printing Office. Wash. D.C. (1960).

37 - DOYLE L.B. -

Indexing and abstracting by association. Am. Doc. Vol. 13. n° 4. October 1967.

38 - DOYLE L.B. -

Is automatic classification a reasonnable application of statistical analysis of text ?
J. ACM Vol. 12. N° 4. October 1965.

39 - DOYLE L.B. -

Semantic road maps for literature searches.
J. ACM Vol. 8. n° 4. October 1961.

40 - EDMUNDSON H.P. -

Problems in automatic abstracting Commun ACM. Vol. 7. n° 4 April 1964.

41 - EDMUNDSON H.P., WYLLYS R.E. -

Automatic abstracting and indexing -
- survey and recommandations Commun ACM. Vol. 4. p. 226. May 1961.

42 - Eléments du Vocabulaire de l'Informatique.

Rev. Informatique Méd. 1 (3). 233-238, septembre 1970;1 (4). 343-348 décembre 1970 ; 2 (1), 67-73. mars 1971.

43 - EMPEY S.L. -

Computer applications in medecine and the biological sciences : Bibliography Commun ACM. Vol. 6, N° 4. pp. 176-183. April 1963.

44 - Excerpta Medica automated storage and retrieval program of biomedical information. The Hague. Excerpta Medica Foundation. 48 pages.

45 - FAIRTHORNE R.A. -

Some basic comments on retrieval testing.
J. Doc. Vol. 24. N° 4. December 196.

46 - First thoughts on MEDLARS The Lancet, pp. 818-819.

April 19, 1969.

47 - FISCHER M. The KWIC index concept. -

A retrospective view Am. Doc. vol. 17. n° 2.
April 1966.

48 - FISKUS M. -

Les problèmes d'économie de codage dans une banque de données accessibles en temps réel (Programme SINBAD).
Journées "Software des Banques de Données "IRIA/AFCET, INSERM, 19 juin 1970.

49 - FORRESTER G. -

Mechanizing a bibliography data bank, FID/IFIP
Conference on Mechanized information Storage, Retrieval and Dissemination. Rome. June 1967.

50 - FREMINVILLE B. de Documentation automatique et aide à la recherche.
Le Concours Médical 22-47,pp 8793-8797, 21 novembre 1970.

51 - GARFIELD E. -

Science citation index.
Science vol. 144. n° 3 619. p. 649. May 8, 1964.

52 - GINGUAY M. -

Dictionnaire d'Informatique. Anglais-Français.
Masson et Cie. Paris 1970. 139 pages.

53 - GORDON B.L. -

Regularization and stylization of medical records.
J.A.M.A. Vol. 212. n° 5. pp. 1 502-1 507. June 1. 1970.

54 - GREMY Y.F., HIREL J.C., MARTIN A.V.J., NAMIAN P. -

Codage de l'information on machine Supplément du Concours Médical n° 16. pp. 15-32, 15 avril 1969.

55 - Guide to MEDLARS Services National Library of Medicine,

Bethesda, Md, 1970, 22 pages. (Public Health Service Publication n° 1694)

56 - HAMBURGER J.

La Médecine. Demain La Presse Médicale 76, n° 6, p 285, 3 février 1968.

57 - HARLEY, A.J. BARRACLOUGH, E.D.

MEDLARS information retrieval in Britain Postgrad Med J 42 : 69-73, February 1966.

58 - HEAULME M. de. -

Introduction à l'exploitation des dossiers médicaux par ordinateur
Cahiers Laënnec pp. 46-59. décembre 1969 - mars 1970.

59 - HENDERSON M.M. -

Bibliography on evaluation of information systems National Bureau of Standards. Washington. D.C., July 1965.

60 - HIBBARD I.N. -

Some combinatorial properties of certain trees with applications to searching and sorting.
J. ACM. Vol. 9, n° 1. pp. 13-28. January 1952.

61 - HOGAN R. -

An evaluation of MEDLARS output : demand and recurring bibliographies.
Bull. Med. Libr. Ass. 54 : 321-324. October 1966.

62 - Information retrieval at W.H.O.

The Lancet p. 333.
August 9, 1969.

63 - Information et recherche médicale.

Le Concours Médical 90-29. pp. 5.565-5.567.
20 juillet 1968.

64 - JOYCE I., NEEDHAM R.M. -

The thesaurus approach to information retrieval.
Am. Doc. Vol. 9, n° 3. July 1958.

65 - KAREL L., AUSTIN, C.J., CUMMINGS M.M. -

Computerized bibliographic services for biomedicine
Science 148 : 766-772. May 9, 1965.

66 - KENT, A. -

Computers and biomedical information storage and retrieval.
J.A.M.A. 196 : 927-932. June 1966.

67 - KENTON C. -

The recurring bibliographies program of MEDLARS.
Bull. Med. Libr. Ass. 54 : 136-137. April 1965.

68 - KESSLER M.M. -

Bibliographic coupling between scientific papers.
Am. Doc. Vol. 14. n° I. pp. 10-25. January 1963.

69 - KESSLER M.M. -

Comparison of the results of bibliographic coupling and analytic subject indexing.
Am. Doc. Vol. 16. n° 3. July 1965.

70 - KIRSCHNER A.K. -

L'informatique et la documentation médicale.
In l'Informatique en Médecine. Association des Médecins de Langue française du Canada, Ottawa pp. 73-80. 1969.

71 - KOEST G., FRANCK C. -

Notions Elémentaires de Bibliographie Médicale. Librairie Le François. Paris 1955. 62 pages.

72 - LACRONIQUE J.F. -

Dossier Médical et ordinateur.
Le Concours Médical 93-23 pp. 4.403-4.407. 5 juin 1971.

73 - LACRONIQUE J.F. -

Quelques applications des techniques de la documentation automatique dans le domaine médical.
Rev. Informatique Med. Vol. 2, n° 2. pp. 161-169. juin 1971.

74 - LANCASTER F.W. -

Evaluation of the performance of a large computerized information system
J.A.M.A. vol 207, n° I pp 114-120, January 6,1969.

75 - LANCASTER F.W. -

Evaluation of the MEDLARS demand search service National Library of Medicine,
Bethesda. Md. 1968.

76 - LANCASTER F.W. -

Information Retrieval Systems.
Characteristics. Testing, and Evaluation.
Wiley. New York 1968. 222 pages.

77 - LANCASTER F.W.

MEDLARS : Rapport on evaluation of its operating efficiency.
Am. Doc. pp 119-142, April 1969.

78 - LATOURNE J.Y. -

La documentation automatique.
Les Cahiers Français n° 147. pp. 76-49. mars-avril 1971.

79 - LEVERY F. -

L'application des machines et des méthodes de traitement automatique de l'information au domaine médical.
La Presse Médicale. 75 n° 7. p. 356. 11 février 1967.

80 - LEVERY F. -

Traduction, documentation, enseignement automatiques.
Science et Vie n° 69. 1964.

81 - LEWIS P.A.W., BAXENDALE P.A., BENNETT J.L. -

Statistical discrimination of the synonym-autonym relationship between words J. ACM. Vol. 14. N° 1.
January 1967.

82 - Lexique Informatique Gazette Médicale de France tome 78.

n° 16. pp. 2.473-2.474. 23 avril 1971.

83 - LIBBEY M.A. -

The use of second order descriptors for document retrieval.
Am. Doc. Vol. 18. n° 1. January 1967.

84 - LICKLIDER J.C. -

Librairies of the Future. The M.I.T. Press.
Cambridge. Mass. Mass. 1965.

85 - LIPETZ B.A. -

Information storage and retrieval.
Sci Am. Vol. 215. n° 3. p. 224. September 1966.

86 - Looking for papers

Brit. Med. J. I : 4,
January 3. 1970.

87 - LOUKOPOULOS L. -

Indexing problems and some of their solutions.
Am. Doc. Vol. 17. n° 1. January 1966.

88 - LUHN H.P. -

The automatic creation of literature abstracts I.B.M.
J. Res Develop. Vol. 2, n° 2. pp. 159-165. April 1956.

89 - LUHN H.P. -

KEYWORD - in - context index for technical literature (K.W.I.C. index);
Am. Doc. II : 288-295. October 1960.

90 - MARGOLIS J. -

Citation indexing and evaluation of scientific papers.
Science vol. 155. n° 3 767. pp. 1 213-1 219. March. 10. 1967.

91 - MARKUS J. -

State of the art of published indexes.
Am. Doc. Vol. 13. n° 1. January 1962.

92 - MARON M.E. -

Automatic indexing : an experimental inquiry.
J. ACM. Vol. 8. n° 3. pp. 404-417. July 1961.

93 - MARON M.E., KUHNS. J.L. -

On relevance, probabilistic indexing, and information retrieval
J. ACM. Vol. 7, n° 3. pp. 216-244. July 1960.

94 - MARTINET, ANDRE. -

Eléments de Linguistique Générale.
Librairie Armand Colin. Paris. 1970. 218 pages.

95 - MASON E.E., BULGREN W.G. -

Computer Applications in Medicine.
Thomas. Springfield 1964. 171 pages.

96 - MASTURZO A. -

Cybernetic Medicine.
Thomas. Springfield. 1965. 142 pages.

97 - La Médecine à l'âge électronique.

Le Concours Médical 89-41. pp. 6.501-6.506.
14 octobre 1967.

98 - Medical Subject Meadings (MeSH). -

Index Medicus vol. II, n° 1, part 2 January 1970. National Library of Medicine.
Washington. U.S. Public Health Service.

99 - MONNEROT-DUMAINE. -

Classement de la littérature médicale mondiale. La réforme de l'Index Medicus.
La Presse Médicale. 71. n° 40. p. 1896. 28 septembre 1963.

100 - MONNEROT-DUMAINE. -

L'Electronique au secours de la documentation médicale.
La Presse Médicale 72. n° 49. p. 2960. 21 novembre 1964.

101 - MOOERS C.N. -

The next twenty years in information retrieval.
Proc. Western Joint Computer Conf. p. 81, March. 1959.

102 - MOOERS C.N.

TRAC - A procedure describing language for the reactive typewriter
Commun ACM. Vol. 9. n° 3. March. 1966.

103 - MOUNIN G. -

Clefs pour la Linguistique. Editions Seghers.
Paris 1971. 188 pages.

104 - NANUS B. -

The use of electronic computers for information retrieval.
Bull. Med. Libr. Ass. 48 : 278-291. July 1960.

105 - NEEDHAM R.M., SPARK-JONES K. -

Keywords and clumps J. Doc. Vol. 20. n° 1. March. 1964.

106 - O'CONNOR J. -

Automatic subject recognition in scientific papers : an empirical study.
J. ACM. Vol. 12. n° 4. pp. 490-515. October 1965.

107 - O'CONNOR J. -

Mechanized indexing methods and their testing.
J. ACM. Vol. II. n° 4. p. 437. October 1964.

108 - OVERHAGE C.F.J. -

Plans for project Intrex.
Science vol. 152. n° 3.725. May 20. 1966.

109 - PIZER I.H. -

A regional medical library network.
Bull. Med. Libr. Ass. 57 : 101-115. April 1969.

110 - Project FAIR. -

J.A.M.A. Vol. 208. nº 7. p. 1.198.
May 19. 1969.

111 - RANNIE I. -

Medical literature retrieval.
Brit. Med. J. 5 499 : 1 351-1 352 May 28. 1966.

112 - RESNICK A. -

Relative effectiveness of document titles and abstracts for determining relevance of documents.
Science vol. 134. nº 3 484. pp. 1 004-1 005. October 6, 1961.

113 - ROCCHIO J.J., SALTON G. -

Information search optimization and iterative retrieval techniques.
Proc. 1965 Fall. Joint Computer Conf. pp. 293-305. 1965.

114 - ROGERS F.B. -

The development of MEDLARS.
Bull. Med. Libr. Ass. 52 : 150-151. January 1964.

115 - ROGERS F.B. -

Stresses in current medical bibliography.
New Eng. J. Med. 267 : 704-707. October 4. 1962.

116 - RUHL M.J. -

Development of a medical specialty recurring bibliography.
Index of Rheumatology.
Bull. Med. Libr. Ass. 55 : 70-74. January 1967.

117 - SABIR aids cancer research

J.A.M.A. vol, 210, nº 11, p 2102, Decembre 15, 1969.

118 - SAGE C.R. -

Comprehensive dissemination of current literature.
Am. Doc. Vol. 17. nº 4. October 1966.

119 - SAGE C.R., ANDERSON R.R., FITZWATER D.R. -

Adoptive information retrieval.
Am. Doc. Vol. 16. nº 3. July 1965.

120 - SALTON G. -

Associative document retrieval techniques using bibliographie information.
J. ACM. Vol. 10. n° 4. p 440. October 1963.

121 - SALTON G. -

Automatic information organization and retrieval.
Mc. Graw-Hill Book. Co., New York. 1968. 514 pages.

122 - SALTON G. -

Automatic text analysis.
Science vol. 168. n° 3 929. pp. 335-343. April 17, 1970.

123 - SALTON G. -

Data manipulation and programming problems in automatic information retrieval.
Commun ACM. Vol. 9. n° 3. March. 1966.

124 - SALTON G. -

The evaluation of automatic retrieval procedures - selected test results using the SMART system.
Am. Doc. Vol. 16. n° 3. pp. 209-222. July 1965.

125 - SALTON G. -

Information processing in Western Europe.
Science vol. 144. n° 3 619. May 8. 1964.

126 - SALTON G. -

Manipulation of tress in information retrieval Commun.
ACM. Vol. 5. n° 2. pp. 103-114. February 1962.

127 - SALTON G. -

Search strategy and the optimization of retrieval effectiveness FID/IFIP.
Conference on Mechanized Information Storage, Retrieval and Dissemination. Rome. June 1967.

128 - SALTON G. -

Some experiments in the generation of word and document associations.
Proc. Fall. Joint Computer Conf. p. 234. 1962.

129 - SALTON G. -

Some hierarchical models for automatic document retrieval.
Am. Doc. Vol. 14. n° 3. July 1963.

130 - SALTON G., LESK M.E. -

The SMART automatic document retrieval system. - an illustration
Commun ACM. Vol. 8. n° 6. pp. 391-398. June 1965.

131 - SAMUELSON K. (Ed.). -

Mechanized Information Storage, Retrieval and Dissemination.
Proc. FID/IFIF. Joint Conference. Rome June 1967. North-Holland
Publishing Co., Amsterdam. 1968.

132 - SELYE H., EMBER G. -

Symbolic Shorthand System (SSS.) for Physiology and Medicine 4 th.
ed. Montréal, Institut de Médecine et de Chirurgie Expérimentales,
Université de Montréal. 1964.

133 - SEWELL W. -

Medical subject headings in MEDLARS.
Bull. Med. Libr. Ass. 52 : 164-170. January 1964.

134 - SHAPIRO P.A., BROSS I.D.J., PRIORE R.E., ANDERSON B.B. -

Information in natural languages.
J.A.M.A. Vol. 207. n° 11. pp. 2080-2084. March. 17. 1969.

135 - SIMMONS R.F. -

Storage and retrieval of aspects of meaning in directed graph structures.
Commun. ACM. Vol. 9. n° 3. March. 1966.

136 - SOLLA PRICE D.J. de. -

Networks of scientific papers.
Science Vol. 149. n° 3683. pp. 510-515 July 30, 1965.

137 - SOUTHERN W.A. -

Mechanized retrieval of bio-medical information.
Meth. Inform. in Med. I : 16-22. January 1962.

138 - STEVENS M.E. -

Automatic indexing : a state of the art-report.
U.S. National Bureau of Standards Monograph. 91. March. 1965.

139 - STILES H.E. -

The association factor in information retrieval.
J. ACM Vol. 8. n° 2. p. 271. April 1961.

140 - SUSSENGUTH E.H. JR. -

The use of tree structures for processing files Commun.
ACM. Vol. 6, n° 5. May 1963.

141 - SWANSON D.R. -

Information retrieval - state of the art.
Proc. Western Joint Computer Conf. p. 239. March. 1961.

142 - SWANSON D.R. -

Searching natural language text by computer.
Science Vol. 132. nº 3 434. pp. 1 099-1 104. October 21. 1960.

143 - SWETS J.A. -

Information retrieval systems Science.
Vol. 141. nº 3 577. July 19. 1963.

144 - Le système MEDLARS à la disposition des chercheurs français

La Presse Médicale 70, nº 4, p 200, 24 janvier 1970.

145 - TAINE S.I. -

Bibliographic aspects of MEDLARS.
Bull. Med. Libr. Ass. 52 : 152-158. January 1964.

146 - TAINE S.I. -

The medical literature analysis and retrieval system.
Bull. Med. Libr. Ass. 51 : 157-167, April 1963.

147 - TALBOTT G.D. -

Hot line to the heart.
J.A.M.A. 196 : 984-966. June 1966.

148 - TAYLOR T.R. -

The Principles of Medical Computing Blackwell Scientific Publications.
Oxford and Edinburg. 1967. 154 pages.

149 - TRYSTRAM J.P. -

La Documentation Automatique.
Dunod. 1971. 124 pages.

150 - VICKERY B.C. -

Techniques Modernes de Documentation. Analyse des Systèmes de Recherche de Documents.
Dunod. Paris. 1962. 170 pages.

151 - WARNET N., DEVILLERS P., LE MINOR M. -

Documentation médicale : un exemple d'application de l'informatique
Rev. Informatique Méd. Vol. 2, nº 2. pp. 171-181. juin 1971.

152 - WEIZENBAUM J. -

Contextual understanding by computers
Commun ACM. Vol. 10. n° 8. August 1967.

153 - WESTBROOK J.H. -

Identifying significant research.
Science Vol. 132. n° 3 435. October 28, 1960.

154 - WOLFF-TERROINE M. -

La documentation et l'informatique médicale.
Gazette Médicale de France tome 78. n° 16. pp. 2.523-2.530. 23 avril 1971.

155 - WOLFF-TERROINE M., DEJAY P. -

Evaluation de l'efficacité d'un système documentaire.
Rev. Informatique. Méd. Vol. 2. pp. 149-160. Juin 1971.

156 - YOUDEN W.W. -

Computer Literature Bibliography 1946 to 1963.
Natl. Bur. Std. (U.S.), Misc. Publ. 266. March. 1965.

157 - ZERAFFA J. -

Application de l'informatique à la documentation médicale.
Cahiers Laënnec pp. 83-113. décembre 1969 - mars 1970.

158 - ZUSMAN F.S., LEDLEY R.C. -

Biomedical information retrieval by means of Tabledex Paper 15 th.
Ann. Conf. on Eng. In Med. and Biol. Chicago. November 1962. p. 22.

II - REFERENCES SUPPLEMENTAIRES

Littérature Récente sur le Projet SMART

A - Livres et Périodiques

1 - SALTON G. -

Automatic Information Organization and Retrieval
McGraw Hill Book Co., New York, 1968, 514 pages.

2 - SALTON, G., LESK, M.E. -

Computer Evaluation of Indexing and Text
Processing, J ACM, vol 15, n° 1, pp. 8-36, January 1968.

3 - SALTON, G. -

Search and Retrieval Experiments in Real Time Information Retrieval,
Preprints of Proceedings IFIP Congress 68, Edinburgh, August 1968.

4 - SALTON, G. -

A Comparison between Manuel and Automatic Indexing Methods
Am. Doc., vol 20, n° 1, January 1969.

5 - SALTON G. (Ed.). -

The SMART Retrieval System Experiment in Automatic Document
Processing. PrenticeHall. 1971.

6 - LESK M.E. SALTON G. -

Relevance Assessments and Retrieval System Evaluation. Information
Storage and Retrieval. Vol 4, n° 4, pp 343-359, December 1968.

7 - LESK, M.E., SALTON G. -

Interactive Search and Retrieval Methods using Automatic Informa-
tion Displays, Proceedings AFIPS Spring Joint Computer Conference,
Thompson Book Co., Washington, pp 435-446, May 1969.

1 – [illegible] [illegible] –

2 – [illegible] [illegible] –

3 – [illegible] [illegible] –

4 – [illegible] [illegible] –

5 – [illegible] [illegible] –

6 – [illegible] [illegible] –

7 – [illegible] [illegible] –

B - I.S R - Scientific Reports to the National Science Foundation Information Storage and Retrieval

1 - Report I S R - 11. -

(Reports on System Design, Evaluation, Feedback and Cluster Searching). Department of Computer Science, Cornell University, June 1966.

2 - Report I S R - 12. -

(Reports on Evaluation, Clustering and Feedback) Dept. Comp. Sci., Cornell Univ, June 1967.

3 - Report I S R - 13. -

(Reports on Evaluation, Procedures and Results 1965-1967) Dept. Comp. Sci., Cornell Univ., December 1967.

4 - Report I S R - 14. -

(Reports on System Design, Analysis, and Feedback) Dept. Comp. Sci., Cornell Univ., October 1968.

5 - Report I S R - 15. -

(Relevance Feedback in an Automatic Document Retrieval System, I D E, E. Master's Thesis). Dept. Comp. Sci, Cornell Univ., January 1969.

6 - Report I S R - 16. -

(Reports on SMART System Design, Content Analysis, User Feedback, and Clustering), Dept, Comp. Sci., Cornell Univ., September 1969.

7 - Report I S R - 17. -

(Document Vector Modification in On-Line Information Retrieval Systems, BRAUEN, T., Master's Thesis) Dept. Comp. Sci., Cornell Univ., September 1969.

oOo

B - I.S.R - Scientific Reports to the National Science Foundation Information Storage and Retrieval

1 - Report I S R - 11. -

(Reports on System Design, Evaluation, Feedback and Cluster Searching). Department of Computer Science, Cornell University, June 1966.

2 - Report I S R - 12. -

(Reports on Evaluation, Clustering and Feedback) Dept. Comp. Sci., Cornell Univ, June 1967.

3 - Report I S R - 13. -

(Reports on Evaluation, Procedures and Results 1965-1967)
Dept. Comp. Sci., Cornell Univ., December 1967.

4 - Report I S R - 14. -

(Reports on System Design, Analysis, and Feedback)
Dept. Comp. Sci., Cornell Univ., October 1968.

5 - Report I S R - 15. -

(Relevance Feedback in an Automatic Document Retrieval System, I D E, E. Master's Thesis). Dept. Comp. Sci, Cornell Univ., January 1969.

6 - Report I S R - 16. -

(Reports on SMART System Design, Content Analysis, User Feedback, and Clustering).
Dept, Comp. Sci., Cornell Univ., September 1969.

7 - Report I S R - 17. -

(Document Vector Modification in On-Line Information Retrieval Systems, HAUSEN, T., Master's Thesis) Dept. Comp. Sci., Cornell Univ., September 1969.

ooo